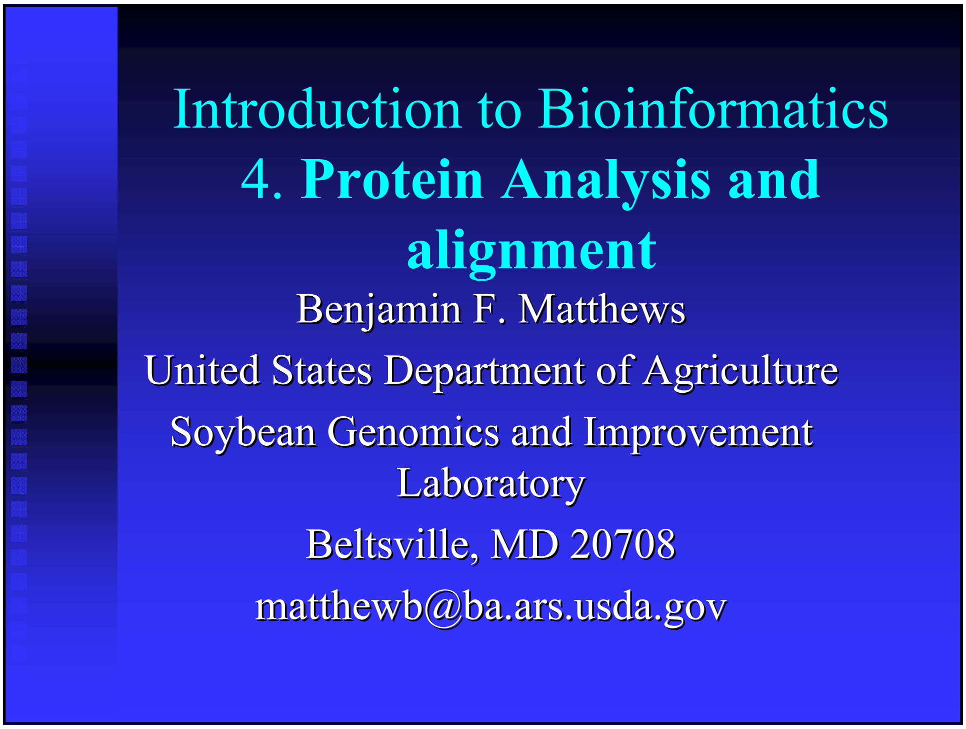


Introduction to Bioinformatics 4. **Protein Analysis and alignment**

Benjamin F. Matthews

United States Department of Agriculture

Soybean Genomics and Improvement Laboratory

Beltsville, MD 20708

matthewb@ba.ars.usda.gov

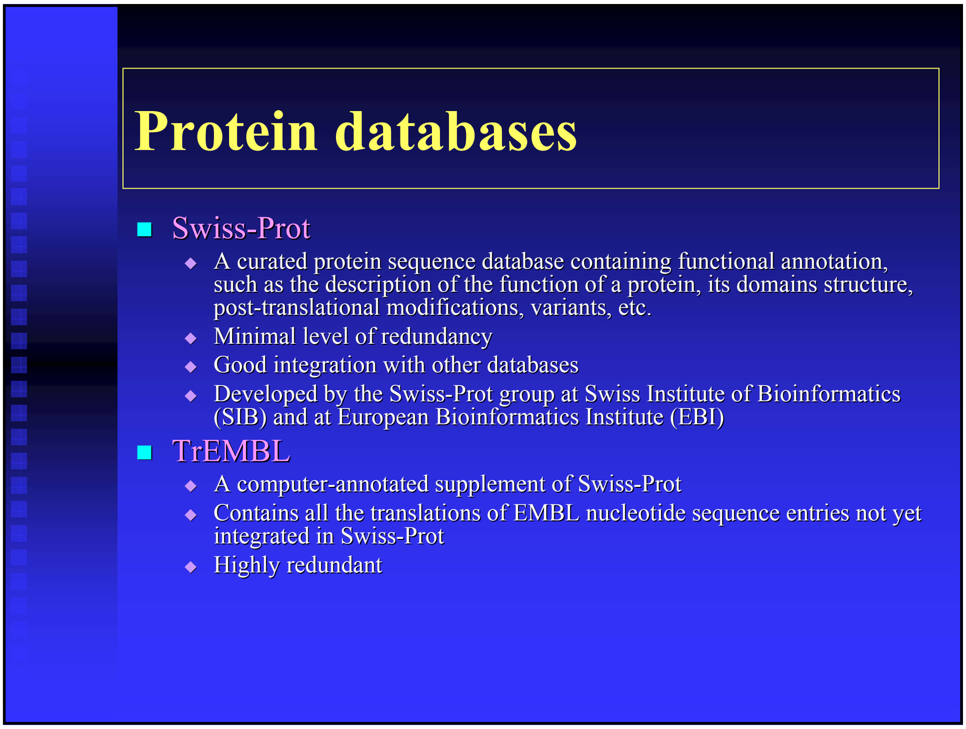
What we will cover today

- DNA translation
  - Protein analysis
- Similarity searches

You obtained the DNA sequence of your cDNA clone

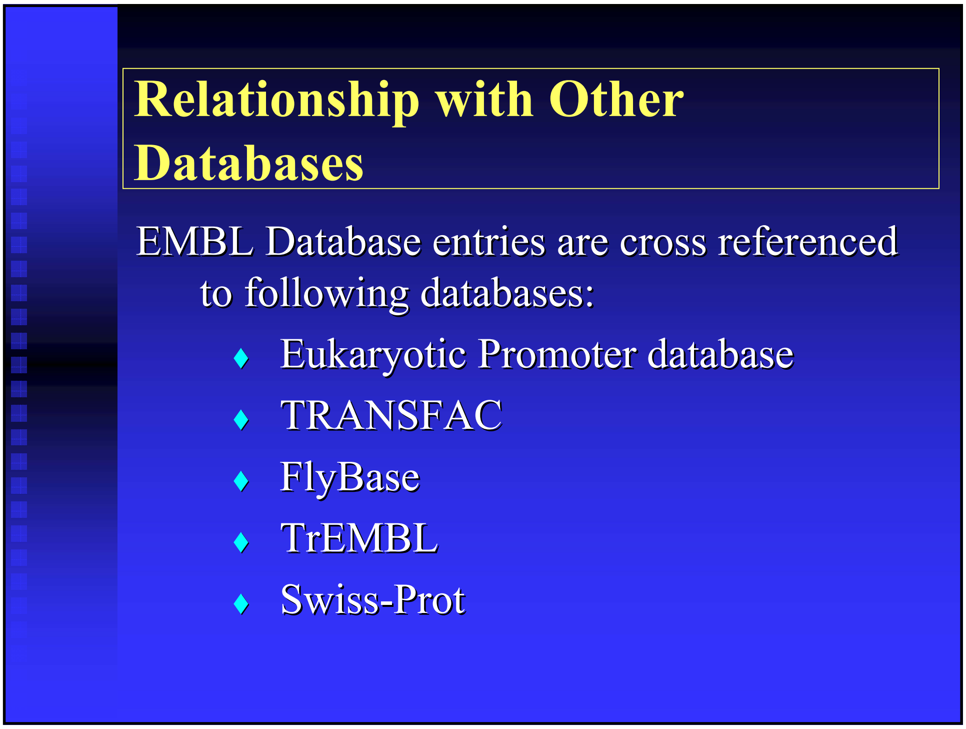
- Does the sequence represent a full-length cDNA?
- What protein does it encode?
- What are the properties of the protein?
- Is the protein amino acid sequence conserved?
- How closely does it resemble proteins of known function?

Translation of DNA sequence into protein sequence



Protein databases

- Swiss-Prot
  - A curated protein sequence database containing functional annotation, such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.
  - Minimal level of redundancy
  - Good integration with other databases
  - Developed by the Swiss-Prot group at Swiss Institute of Bioinformatics (SIB) and at European Bioinformatics Institute (EBI)
- TrEMBL
  - A computer-annotated supplement of Swiss-Prot
  - Contains all the translations of EMBL nucleotide sequence entries not yet integrated in Swiss-Prot
  - Highly redundant



Relationship with Other Databases

EMBL Database entries are cross referenced to following databases:

- Eukaryotic Promoter database
- TRANSFAC
- FlyBase
- TrEMBL
- Swiss-Prot

ExPASy

- **<u>Ex</u>**pert **<u>P</u>**rotein **<u>A</u>**nalysis **<u>Sy</u>**stem
- Swiss Institute of Bioinformatics
- Proteomics server for protein analysis
- http://us.expasy.org/ - in US
- **http://www.expasy.org/** -in Switzerland
- Translate tool, other tools, molecular databases, and links

ExPaSy Databases

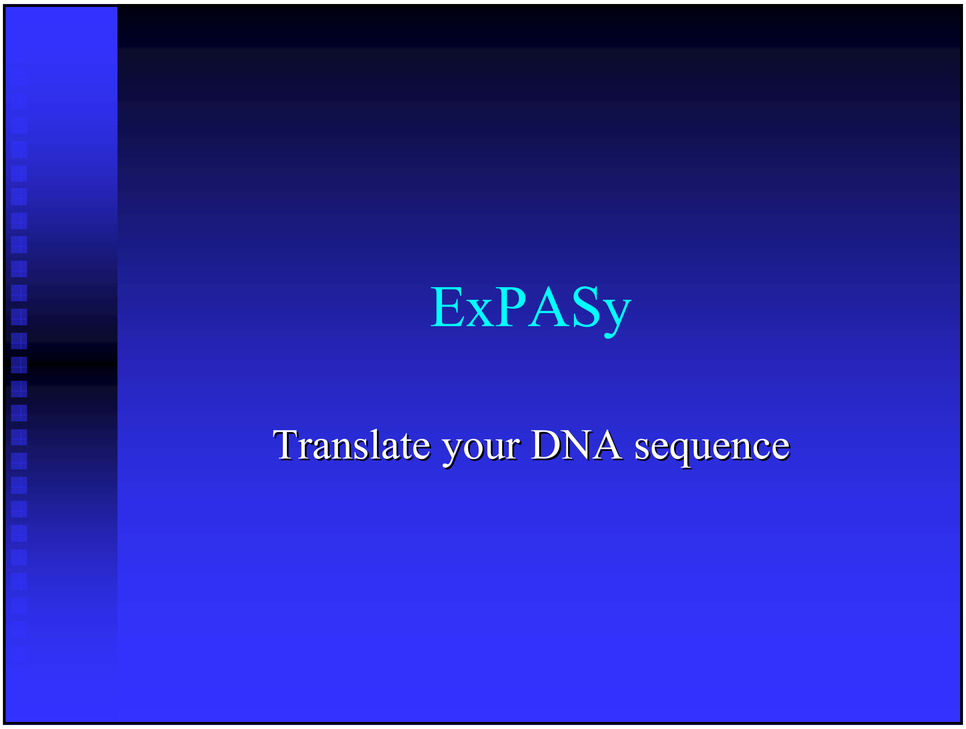
Swiss-Prot:	protein database
TrEMBL:	protein database
Prosite:	protein families and domains
Swiss-2Dpage:	2D polyacrylamide gel electrophoresis
Swiss-3Dimage:	3D images of proteins and other biological macromolecules
Enzyme:	enzyme nomenclature
CD40Lbase:	CD40 ligand defects
SeqAnalRef:	sequence analysis bibliographic references

ExPaSy Tools

- http://bo.expasy.org/
- Protein and sequence analysis tools
- Melanie 4 - Software for 2-D PAGE analysis
- Roche Applied Science's Biochemical Pathways

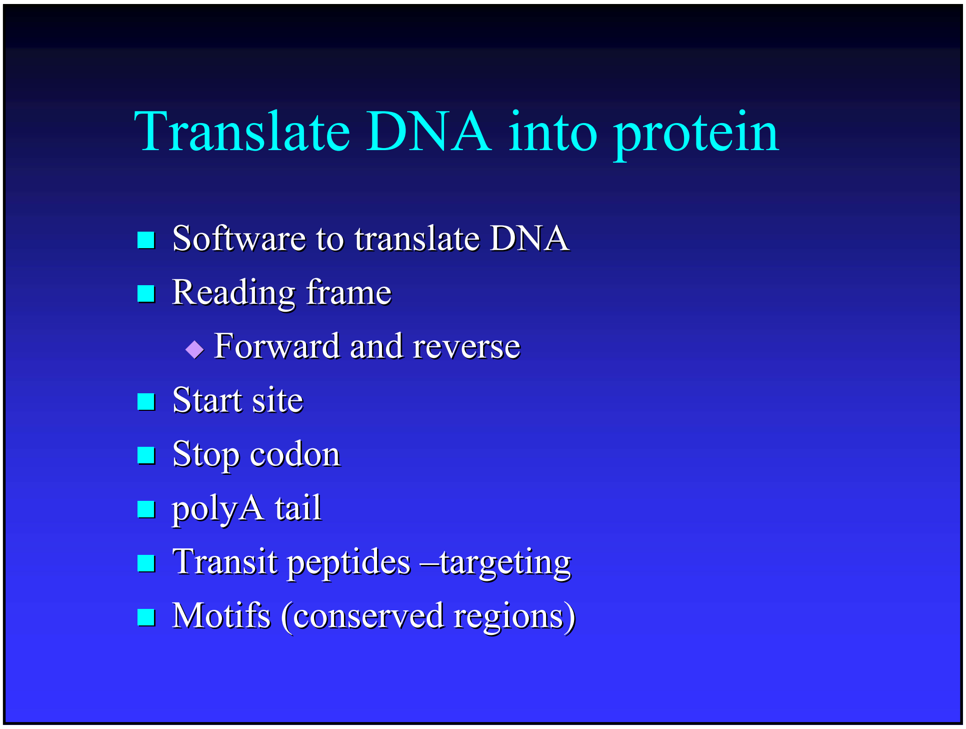
Ensemble

- http://www.ensembl.org/
- A joint project between EMBL-EBI and the Sanger Institute to develop a software system produces and maintains automatic annotation on eukaryotic genomes.



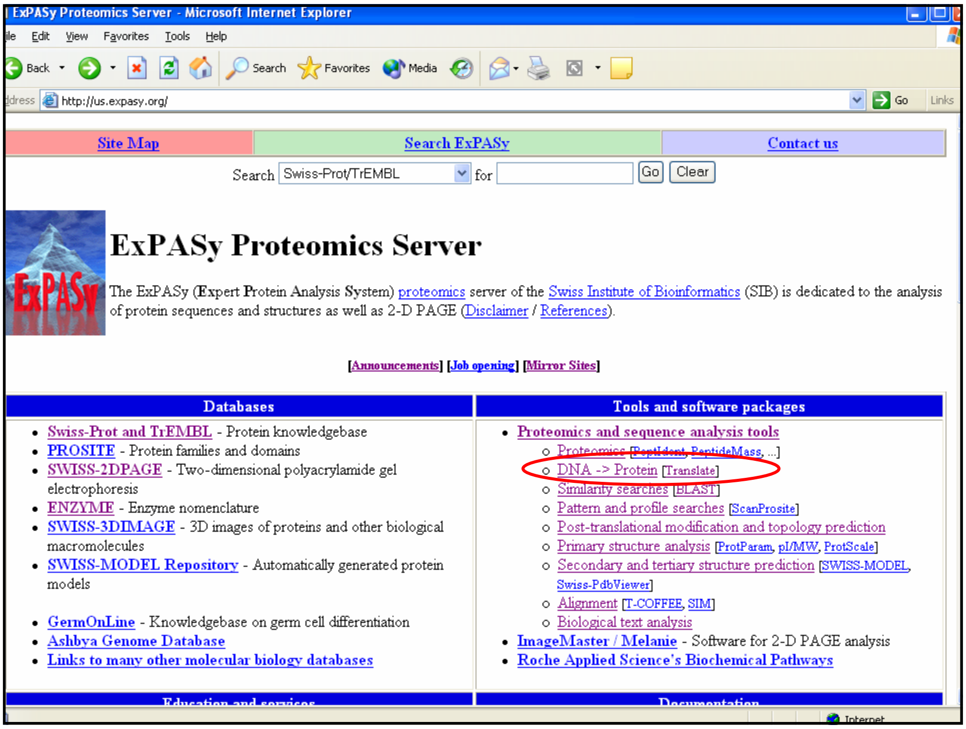
ExPASy

Translate your DNA sequence



Translate DNA into protein

- Software to translate DNA
- Reading frame
  - Forward and reverse
- Start site
- Stop codon
- polyA tail
- Transit peptides –targeting
- Motifs (conserved regions)


ExPASy Proteomics Server - Microsoft Internet Explorer
ile Edit View Favorites Tools Help
Back
Search
Favorites
Media
ddress
Go
Links
Internet

http://us.expasy.org/

Site Map | Search ExPASy | Contact us

Search Swiss-Prot/TrEMBL for Go Clear

ExPASy Proteomics Server

The ExPASy (Expert Protein Analysis System) proteomics server of the Swiss Institute of Bioinformatics (SIB) is dedicated to the analysis of protein sequences and structures as well as 2-D PAGE (Disclaimer / References).

[Announcements] [Job opening] [Mirror Sites]

Databases

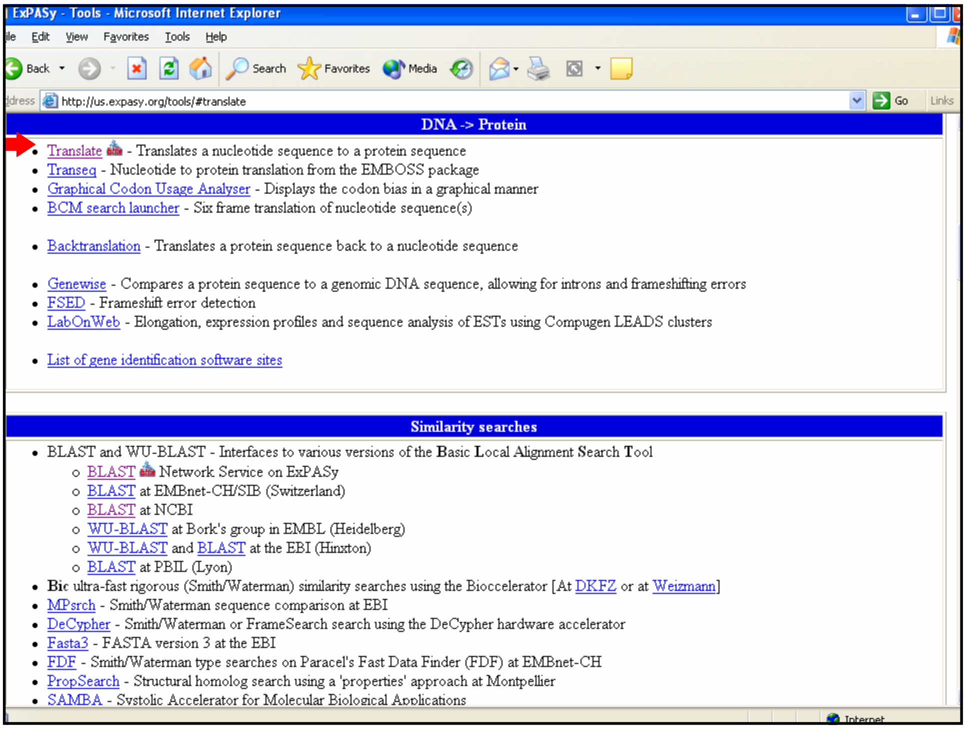
- Swiss-Prot and TrEMBL - Protein knowledgebase
- PROSITE - Protein families and domains
- SWISS-2DPAGE - Two-dimensional polyacrylamide gel electrophoresis
- ENZYME - Enzyme nomenclature
- SWISS-3DIMAGE - 3D images of proteins and other biological macromolecules
- SWISS-MODEL Repository - Automatically generated protein models

- GermOnLine - Knowledgebase on germ cell differentiation
- Ashbya Genome Database
- Links to many other molecular biology databases

Tools and software packages

- Proteomics and sequence analysis tools
  - Proteomics [PeptIdent, PeptideMass, ...]
  - DNA -> Protein [Translate]
  - Similarity searches [BLAST]
  - Pattern and profile searches [ScanProsite]
  - Post-translational modification and topology prediction
  - Primary structure analysis [ProtParam, pI/MW, ProtScale]
  - Secondary and tertiary structure prediction [SWISS-MODEL, Swiss-PdbViewer]
  - Alignment [T-COFFEE, SIM]
  - Biological text analysis
- ImageMaster / Melanie - Software for 2-D PAGE analysis
- Roche Applied Science's Biochemical Pathways

Education and services | Documentation


ExPASy - Tools - Microsoft Internet Explorer
ile Edit View Favorites Tools Help
Back
Search
Favorites
Media
ddress
Go
Links
Internet

http://us.expasy.org/tools/#translate

DNA -> Protein

- Translate - Translates a nucleotide sequence to a protein sequence
- Transeq - Nucleotide to protein translation from the EMBOSS package
- Graphical Codon Usage Analyser - Displays the codon bias in a graphical manner
- BCM search launcher - Six frame translation of nucleotide sequence(s)

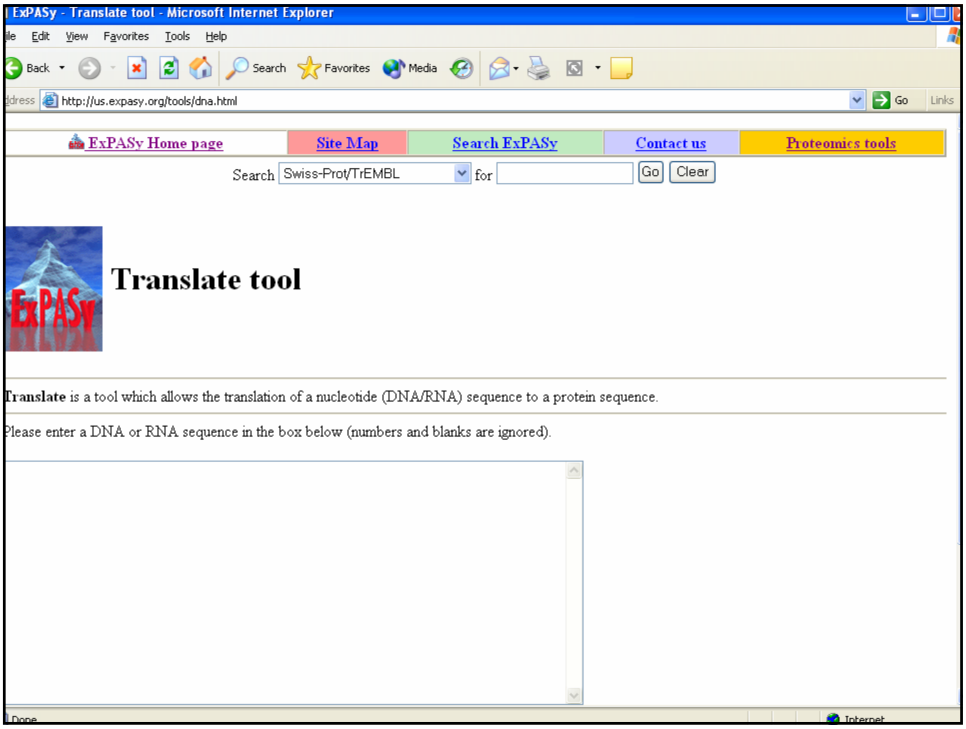
- Backtranslation - Translates a protein sequence back to a nucleotide sequence

- Genewise - Compares a protein sequence to a genomic DNA sequence, allowing for introns and frameshifting errors
- FSED - Frameshift error detection
- LabOnWeb - Elongation, expression profiles and sequence analysis of ESTs using Compugen LEADS clusters

- List of gene identification software sites

Similarity searches

- BLAST and WU-BLAST - Interfaces to various versions of the Basic Local Alignment Search Tool
  - BLAST Network Service on ExPASy
  - BLAST at EMBnet-CH/SIB (Switzerland)
  - BLAST at NCBI
  - WU-BLAST at Bork's group in EMBL (Heidelberg)
  - WU-BLAST and BLAST at the EBI (Hinxton)
  - BLAST at PBIL (Lyon)
- Bic ultra-fast rigorous (Smith/Waterman) similarity searches using the Bioccelerator [At DKFZ or at Weizmann]
- MPsrch - Smith/Waterman sequence comparison at EBI
- DeCypher - Smith/Waterman or FrameSearch search using the DeCypher hardware accelerator
- Fasta3 - FASTA version 3 at the EBI
- FDF - Smith/Waterman type searches on Paracel's Fast Data Finder (FDF) at EMBnet-CH
- PropSearch - Structural homolog search using a 'properties' approach at Montpellier
- SAMBA - Systolic Accelerator for Molecular Biological Applications

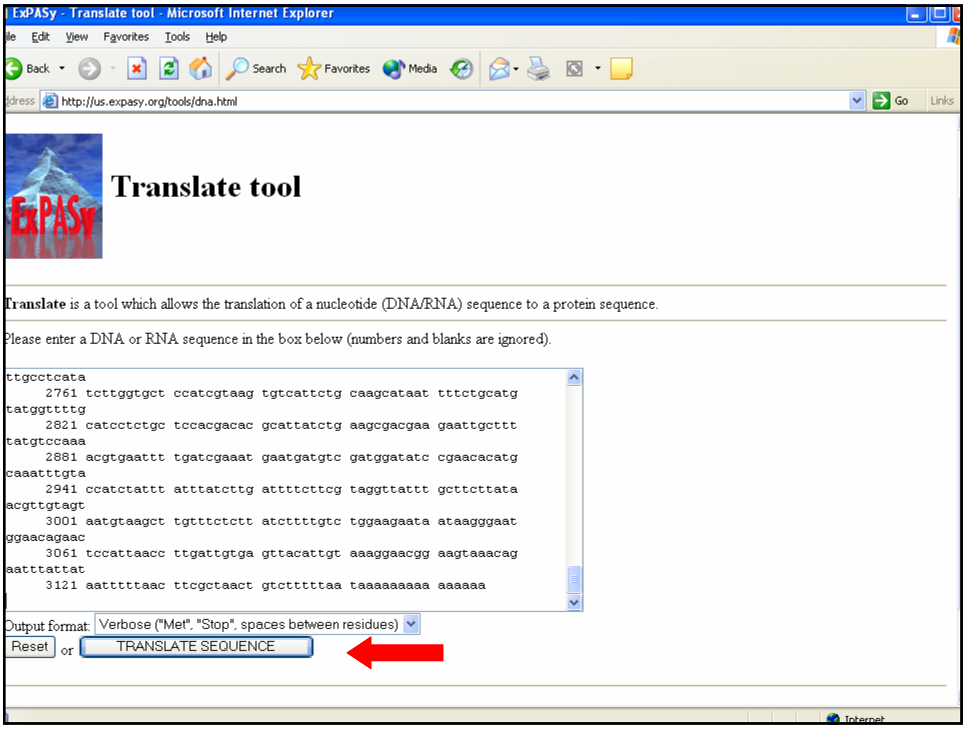


ExPASy - Translate tool - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Media
Address http://us.expasy.org/tools/dna.html
Go
Links
ExPASy Home page
Site Map
Search ExPASy
Contact us
Proteomics tools
Search Swiss-Prot/TrEMBL for
Go
Clear
Done
Internet

Translate tool

Translate is a tool which allows the translation of a nucleotide (DNA/RNA) sequence to a protein sequence.

Please enter a DNA or RNA sequence in the box below (numbers and blanks are ignored).



ExPASy - Translate tool - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Media
Address http://us.expasy.org/tools/dna.html
Go
Links
Internet

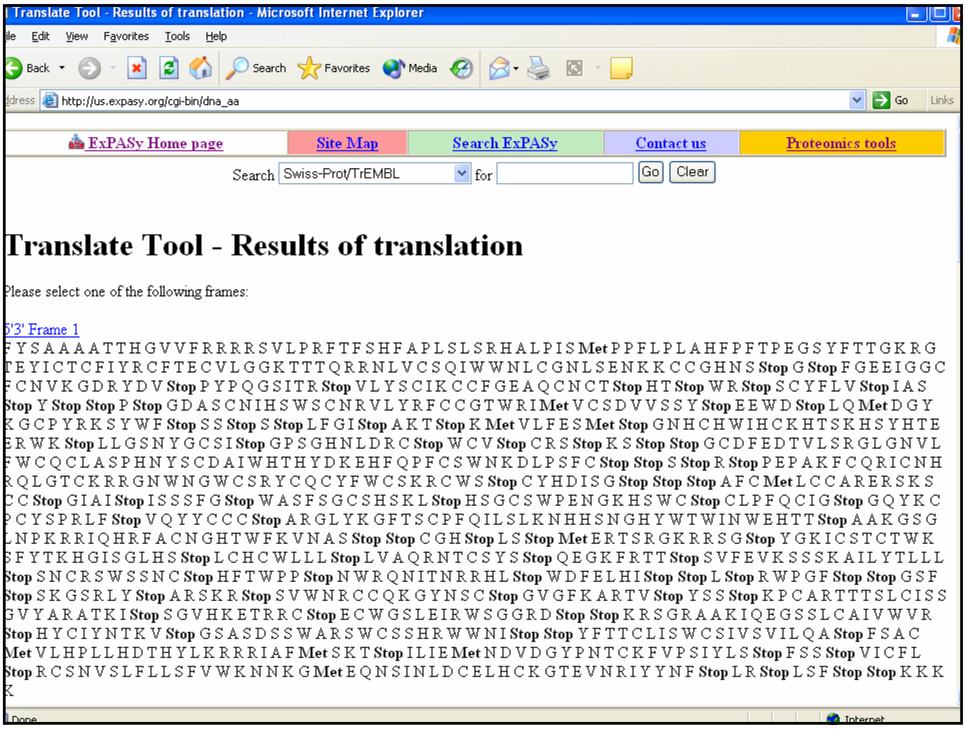
Translate tool

Translate is a tool which allows the translation of a nucleotide (DNA/RNA) sequence to a protein sequence.

Please enter a DNA or RNA sequence in the box below (numbers and blanks are ignored).

```
ttgcctcata
     2761 tcttggtgct ccatcgtaag tgtcattctg caagcataat tttctgcatg
tatggttttg
     2821 catcctctgc tccacgacac gcattatctg aagcgacgaa gaattgcttt
tatgtccaaa
     2881 acgtgaattt tgatcgaaat gaatgatgtc gatggatatc cgaacacatg
caaatttgta
     2941 ccatctattt atttatcttg attttcttcg taggttattt gcttcttata
acgttgtagt
     3001 aatgtaagct tgtttctctt atcttttgtc tggaagaata ataagggaat
ggaacagaac
     3061 tccattaacc ttgattgtga gttacattgt aaaggaacgg aagtaaacag
aatttattat
     3121 aatttttaac ttcgctaact gtctttttaa taaaaaaaaa aaaaaa
```

Output format: Verbose ("Met", "Stop", spaces between residues)



ddress http://us.expasy.org/cgi-bin/dna_aa

Search Swiss-Prot/TrEMBL for

Translate Tool - Results of translation

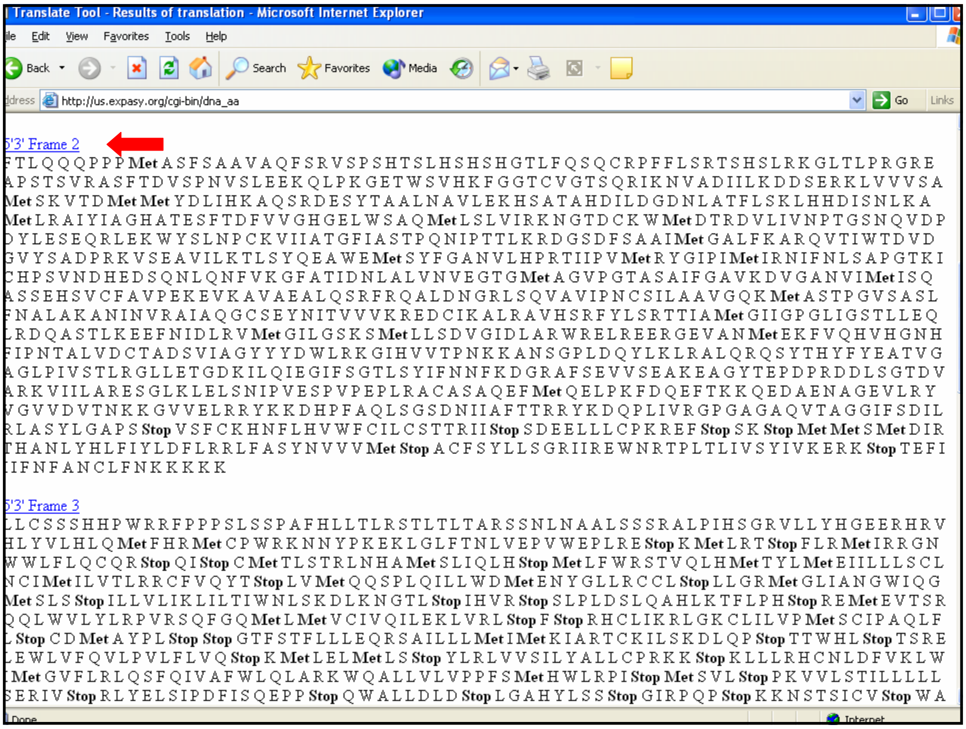
Please select one of the following frames:

5'3' Frame 1

```
F Y S A A A A T T H G V V F R R R R S V L P R F T F S H F A P L S L S R H A L P I S Met P P F L P L A H F P F T P E G S Y F T T G K R G
T E Y I C T C F I Y R C F T E C V L G G K T T T Q R R N L V C S Q I W W N L C G N L S E N K K C C G H N S Stop G Stop F G E E I G G C
F C N V K G D R Y D V Stop P Y P Q G S I T R Stop V L Y S C I K C C F G E A Q C N C T Stop H T Stop W R Stop S C Y F L V Stop I A S
Stop Y Stop Stop P Stop G D A S C N I H S W S C N R V L Y R F C C G T W R I Met V C S D V V S S Y Stop E E W D Stop L Q Met D G Y
K G C P Y R K S Y W F Stop S S Stop S Stop L F G I Stop A K T Stop K Met V L F E S Met Stop G N H C H W I H C K H T S K H S Y H T E
E R W K Stop L L G S N Y G C S I Stop G P S G H N L D R C Stop W C V Stop C R S Stop K S Stop Stop G C D F E D T V L S R G L G N V L
F W C Q C L A S P H N Y S C D A I W H T H Y D K E H F Q P F C S W N K D L P S F C Stop Stop S Stop R Stop P E P A K F C Q R I C N H
R Q L G T C K R R G N W N G W C S R Y C Q C Y F W C S K R C W S Stop C Y H D I S G Stop Stop Stop A F C Met L C C A R E R S K S
C C Stop G I A I Stop I S S S F G Stop W A S F S G C S H S K L Stop H S G C S W P E N G K H S W C Stop C L P F Q C I G Stop G Q Y K C
P C Y S P R L F Stop V Q Y Y C C C Stop A R G L Y K G F T S C P F Q I L S L K N H H S N G H Y W T W I N W E H T T Stop A A K G S G
L N P K R R I Q H R F A C N G H T W F K V N A S Stop Stop C G H Stop L S Stop Met E R T S R G K R R S G Stop Y G K I C S T C T W K
S F Y T K H G I S G L H S Stop L C H C W L L L Stop L V A Q R N T C S Y S Stop Q E G K F R T T Stop S V F E V K S S S K A I L Y T L L L
Stop S N C R S W S S N C Stop H F T W P P Stop N W R Q N I T N R R H L Stop W D F E L H I Stop Stop L Stop R W P G F Stop Stop G S F
Stop S K G S R L Y Stop A R S K R Stop S V W N R C C Q K G Y N S C Stop G V G F K A R T V Stop Y S S Stop K P C A R T T T S L C I S S
G V Y A R A T K I Stop S G V H K E T R R C Stop E C W G S L E I R W S G G R D Stop Stop K R S G R A A K I Q E G S S L C A I V W V R
Stop H Y C I Y N T K V Stop G S A S D S S W A R S W C S S H R W W N I Stop Stop Y F T T C L I S W C S I V S V I L Q A Stop F S A C
Met V L H P L L H D T H Y L K R R R I A F Met S K T Stop I L I E Met N D V D G Y P N T C K F V P S I Y L S Stop F S S Stop V I C F L
Stop R C S N V S L F L L S F V W K N N K G Met E Q N S I N L D C E L H C K G T E V N R I Y Y N F Stop L R Stop L S F Stop Stop K K K
K
```

Results

- Six reading frames provided
- Select one
- Clues:
  - Number and placement of stop codons
  - ATG start site (methionine)
  - Poly (A) tail
  - Alignment with other protein sequences


Translate Tool - Results of translation - Microsoft Internet Explorer
ile Edit View Favorites Tools Help
Back
Search
Favorites
Media
ddress
Go
Links

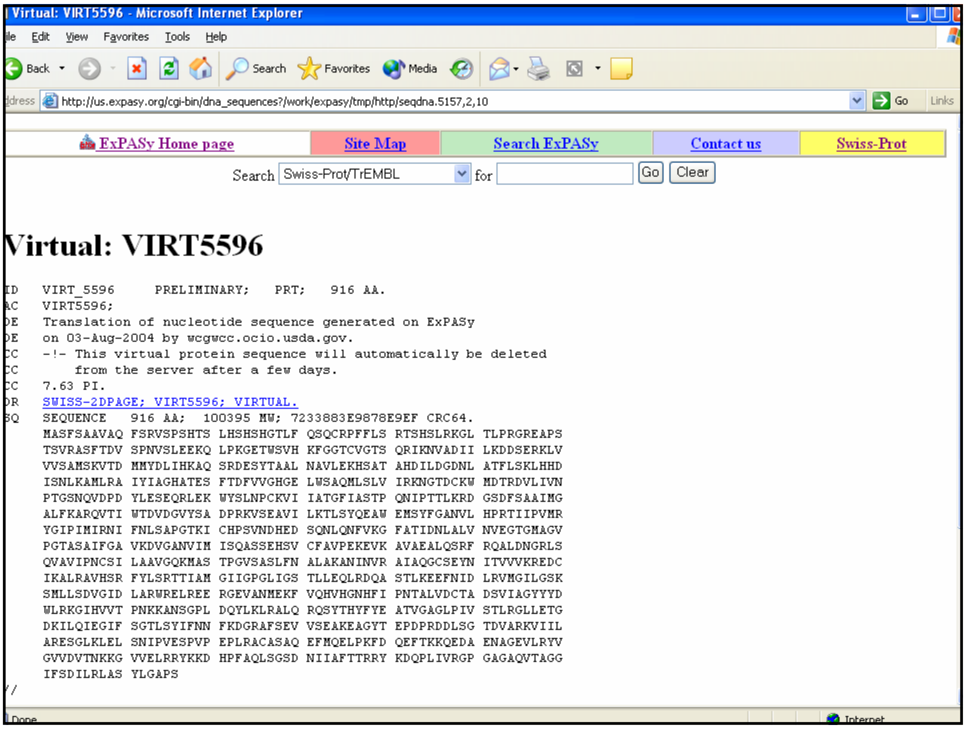
http://us.expasy.org/cgi-bin/dna_aa

5'3' Frame 2

```
FTLQQQPPP Met ASFSAAVAQFSRVSPSHTSLHSHSHGTLFQSQCRPFFLSRTSHSLRKGLTLPRGRE
APSTSVRASFTDVSPNVSLEEKQLPKGETWSVHKFGGTCVGTSQRIKNVADIILKDDSERKLVVVSA
Met SKVTD Met Met YDLIHKAQSRDESYTAALNAVLEKHSATAHDILDGDNLATFLSKLHHDISNLKA
Met LRAIYIAGHATESFTDFVVGHGELWSAQ Met LSLVIRKNGTDCKW Met DTRDVLIVNPTGSNQVDP
DYLESEQRLEKWYSLNPCKVIIATGFIASTPQNIPTTLKRDGSDFSAAI Met GALFKARQVTIWTDVD
GVYSADPRKVSEAVILKTLSYQEAWE Met SYFGANVLHPRTIIPV Met RYGIPI Met IRNIFNLSAPGTKI
CHPSVNDHEDSQNLQNFVKGFATIDNLALVNVEGTG Met AGVPGTASAIFGAVKDVGANVI Met ISQ
ASSEHSVCFAVPEKEVKAVAEALQSRFRQALDNGRLSQVAVIPNCSILAAVGQK Met ASTPGVSASL
FNALAKANINVRAIAQGCSEYNITVVVKREDCIKALRAVHSRFYLSRTTIA Met GIIGPGLIGSTLLEQ
LRDQASTLKEEFNIDLRV Met GILGSKS Met LLSDVGIDLARWRELREERGEVAN Met EKFVQHVHGNH
FIPNTALVDCTADSVIAGYYYDWLRKGIHVVTPNKKANSGPLDQYLKLRALQRQSYTHYFYEATVG
AGLPIVSTLRGLLETGDKILQIEGIFSGTLSYIFNNFKDGRAFSEVVSEAKEAGYTEPDPRDDLSGTDV
ARKVIILARESGLKLELSNIPVESPVPEPLRACASAQEF Met QELPKFDQEFTKKQEDAENAGEVLRY
VGVVDVTNKKGVVELRRYKKDHPFAQLSGSDNIIAFTTRRYKDQPLIVRGPGAGAQVTAGGIFSDIL
RLASYLGAPS Stop VSFCKHNFLHVWFCILCSTTRII Stop SDEELLLCPKREF Stop SK Stop Met Met S Met DIR
THANLYHLFIYLDFLRRLFASYNVVV Met Stop ACFSYLLSGRIIREWNRTPLTLIVSYIVKERK Stop TEFI
IIFNFANCLFNKKKKK
```

5'3' Frame 3

```
LLCSSSHHPWRRFPPPSLSSPAFHLLTLRSTLTLTARSSNLNAALSSSRALPIHSGRVLLYHGEERHRV
HLYVLHLQ Met FHR Met CPWRKNNYPKEKLGLFTNLVEPVWEPLRE Stop K Met LRT Stop FLR Met IRRGN
WWLFLQCQR Stop QI Stop C Met TLSTRLNHA Met SLIQLH Stop Met LFWRSTVQLH Met TYL Met EIILLLSCL
NCI Met ILVTLRRCFVQYT Stop LV Met QQSPLQILLWD Met ENYGLLRCCL Stop LLGR Met GLIANGWIQG
Met SLS Stop ILLVLIKLILTIWNLSKDLKNGTL Stop IHVR Stop SLPLDSLQAHLKTFLPH Stop RE Met EVTSR
QQLWVLYLRPVRSQFGQ Met L Met VCIVQILEKLVRL Stop F Stop RHCLIKRLGKCLILVP Met SCIPAQLF
L Stop CD Met AYPL Stop Stop GTFSTFLLLEQRSAILLL Met I Met KIARTCKILSKDLQP Stop TTWHL Stop TSRE
LEWLVFQVLPVLFLVQ Stop K Met LEL Met LS Stop YLRLVVSILYALLCPRKK Stop KLLLRHCNLDFVKLW
I Met GVFLRLQSFQIVAFWLQLARKWQALLVLVPPFS Met HWLRPI Stop Met SVL Stop PKVVLSTILLLLL
SERIV Stop RLYELSIPDFISQEPP Stop QWALLDLD Stop LGAHYLSS Stop GIRPQP Stop KKNSTSICV Stop WA
```

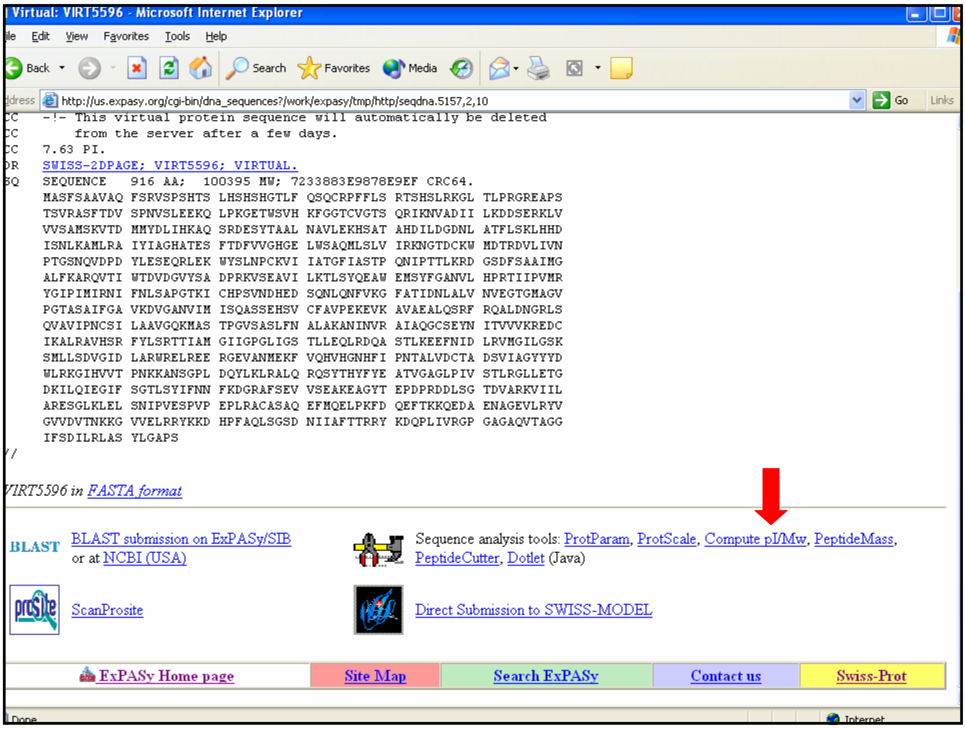


http://us.expasy.org/cgi-bin/dna_sequences?/work/expasy/tmp/http/seqdna.5157,2,10

Swiss-Prot/TrEMBL

Virtual: VIRT5596

```
ID   VIRT_5596     PRELIMINARY;   PRT;   916 AA.
AC   VIRT5596;
DE   Translation of nucleotide sequence generated on ExPASy
DE   on 03-Aug-2004 by wcgwcc.ocio.usda.gov.
CC   -!- This virtual protein sequence will automatically be deleted
CC       from the server after a few days.
CC   7.63 PI.
DR   SWISS-2DPAGE; VIRT5596; VIRTUAL.
SQ   SEQUENCE   916 AA;  100395 MW; 7233883E9878E9EF CRC64.
     MASFSAAVAQ FSRVSPSHTS LHSHSHGTLF QSQCRPFFLS RTSHSLRKGL TLPRGREAPS
     TSVRASFTDV SPNVSLEEKQ LPKGETWSVH KFGGTCVGTS QRIKNVADII LKDDSERKLV
     VVSAMSKVTD MMYDLIHKAQ SRDESYTAAL NAVLEKHSAT AHDILDGDNL ATFLSKLHHD
     ISNLKAMLRA IYIAGHATES FTDFVVGHGE LWSAQMLSLV IRKNGTDCKW MDTRDVLIVN
     PTGSNQVDPD YLESEQRLEK WYSLNPCKVI IATGFIASTP QNIPTTLKRD GSDFSAAIMG
     ALFKARQVTI WTDVDGVYSA DPRKVSEAVI LKTLSYQEAW EMSYFGANVL HPRTIIPVMR
     YGIPIMIRNI FNLSAPGTKI CHPSVNDHED SQNLQNFVKG FATIDNLALV NVEGTGMAGV
     PGTASAIFGA VKDVGANVIM ISQASSEHSV CFAVPEKEVK AVAEALQSRF RQALDNGRLS
     QVAVIPNCSI LAAVGQKMAS TPGVSASLFN ALAKANINVR AIAQGCSEYN ITVVVKREDC
     IKALRAVHSR FYLSRTTIAM GIIGPGLIGS TLLEQLRDQA STLKEEFNID LRVMGILGSK
     SMLLSDVGID LARWRELREE RGEVANMEKF VQHVHGNHFI PNTALVDCTA DSVIAGYYYD
     WLRKGIHVVT PNKKANSGPL DQYLKLRALQ RQSYTHYFYE ATVGAGLPIV STLRGLLETG
     DKILQIEGIF SGTLSYIFNN FKDGRAFSEV VSEAKEAGYT EPDPRDDLSG TDVARKVIIL
     ARESGLKLEL SNIPVESPVP EPLRACASAQ EFMQELPKFD QEFTKKQEDA ENAGEVLRYV
     GVVDVTNKKG VVELRRYKKD HPFAQLSGSD NIIAFTTRRY KDQPLIVRGP GAGAQVTAGG
     IFSDILRLAS YLGAPS
//
```



Virtual: VIRT5596 - Microsoft Internet Explorer
ile Edit View Favorites Tools Help
Back
Search
Favorites
Media
ddress http://us.expasy.org/cgi-bin/dna_sequences?/work/expasy/tmp/http/seqdna.5157,2,10
Go
Links

```
CC   -!- This virtual protein sequence will automatically be deleted
         from the server after a few days.
CC   7.63 PI.
DR   SWISS-2DPAGE; VIRT5596; VIRTUAL.
SQ   SEQUENCE   916 AA;  100395 MW;  7233883E9878E9EF CRC64.
     MASFSAAVAQ FSRVSPSHTS LHSHSHGTLF QSQCRPFFLS RTSHSLRKGL TLPRGREAPS
     TSVRASFTDV SPNVSLEEKQ LPKGETWSVH KFGGTCVGTS QRIKNVADII LKDDSERKLV
     VVSAMSKVTD MMYDLIHKAQ SRDESYTAAL NAVLEKHSAT AHDILDGDNL ATFLSKLHHD
     ISNLKAMLRA IYIAGHATES FTDFVVGHGE LWSAQMLSLV IRKNGTDCKW MDTRDVLIVN
     PTGSNQVDPD YLESEQRLEK WYSLNPCKVI IATGFIASTP QNIPTTLKRD GSDFSAAIMG
     ALFKARQVTI WTDVDGVYSA DPRKVSEAVI LKTLSYQEAW EMSYFGANVL HPRTIIPVMR
     YGIPIMIRNI FNLSAPGTKI CHPSVNDHED SQNLQNFVKG FATIDNLALV NVEGTGMAGV
     PGTASAIFGA VKDVGANVIM ISQASSEHSV CFAVPEKEVK AVAEALQSRF RQALDNGRLS
     QVAVIPNCSI LAAVGQKMAS TPGVSASLFN ALAKANINVR AIAQGCSEYN ITVVVKREDC
     IKALRAVHSR FYLSRTTIAM GIIGPGLIGS TLLEQLRDQA STLKEEFNID LRVMGILGSK
     SMLLSDVGID LARWRELREE RGEVANMEKF VQHVHGNHFI PNTALVDCTA DSVIAGYYYD
     WLRKGIHVVT PNKKANSGPL DQYLKLRALQ RQSYTHYFYE ATVGAGLPIV STLRGLLETG
     DKILQIEGIF SGTLSYIFNN FKDGRAFSEV VSEAKEAGYT EPDPRDDLSG TDVARKVIIL
     ARESGLKLEL SNIPVESPVP EPLRACASAQ EFMQELPKFD QEFTKKQEDA ENAGEVLRYV
     GVVDVTNKKG VVELRRYKKD HPFAQLSGSD NIIAFTTRRY KDQPLIVRGP GAGAQVTAGG
     IFSDILRLAS YLGAPS
//
```

VIRT5596 in FASTA format

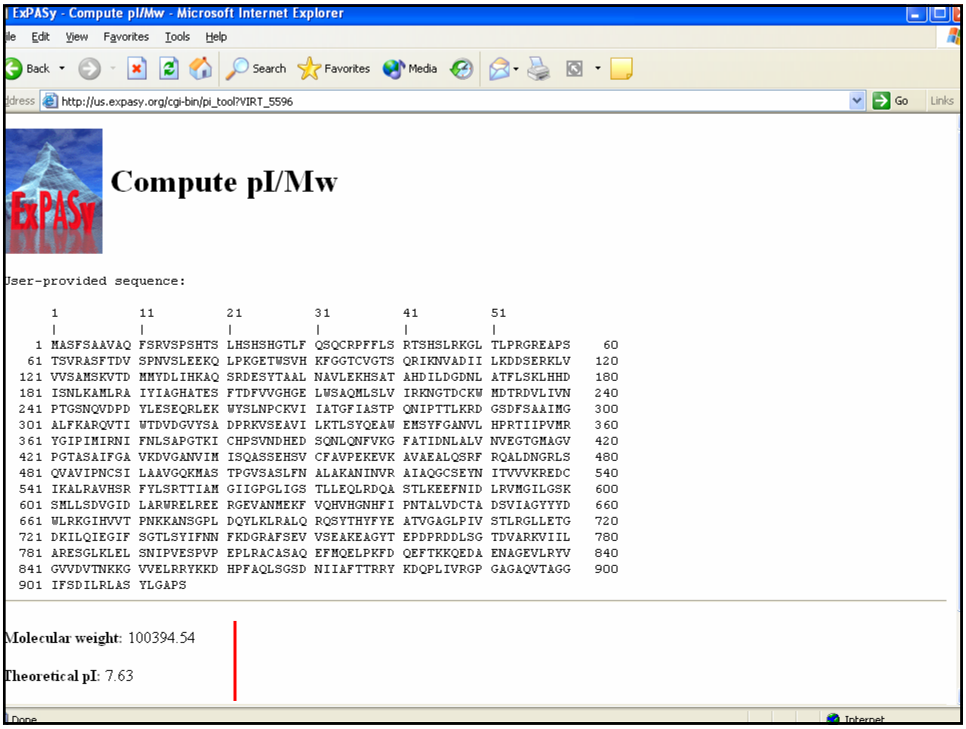
BLAST submission on ExPASy/SIB or at NCBI (USA)

Sequence analysis tools: ProtParam, ProtScale, Compute pI/Mw, PeptideMass, PeptideCutter, Dotlet (Java)

ScanProsite

Direct Submission to SWISS-MODEL

BLAST
ExPASy Home page
Site Map
Search ExPASy
Contact us
Swiss-Prot
Done
Internet



ExPASy - Compute pI/Mw - Microsoft Internet Explorer
ile Edit View Favorites Tools Help
Back
Search
Favorites
Media
ddress http://us.expasy.org/cgi-bin/pi_tool?VIRT_5596
Go
Links
ExPASy

Compute pI/Mw

User-provided sequence:

```
     1          11         21         31         41         51
     |          |          |          |          |          |
   1 MASFSAAVAQ FSRVSPSHTS LHSHSHGTLF QSQCRPFFLS RTSHSLRKGL TLPRGREAPS    60
  61 TSVRASFTDV SPNVSLEEKQ LPKGETWSVH KFGGTCVGTS QRIKNVADII LKDDSERKLV   120
 121 VVSAMSKVTD MMYDLIHKAQ SRDESYTAAL NAVLEKHSAT AHDILDGDNL ATFLSKLHHD   180
 181 ISNLKAMLRA IYIAGHATES FTDFVVGHGE LWSAQMLSLV IRKNGTDCKW MDTRDVLIVN   240
 241 PTGSNQVDPD YLESEQRLEK WYSLNPCKVI IATGFIASTP QNIPTTLKRD GSDFSAAIMG   300
 301 ALFKARQVTI WTDVDGVYSA DPRKVSEAVI LKTLSYQEAW EMSYFGANVL HPRTIIPVMR   360
 361 YGIPIMIRNI FNLSAPGTKI CHPSVNDHED SQNLQNFVKG FATIDNLALV NVEGTGMAGV   420
 421 PGTASAIFGA VKDVGANVIM ISQASSEHSV CFAVPEKEVK AVAEALQSRF RQALDNGRLS   480
 481 QVAVIPNCSI LAAVGQKMAS TPGVSASLFN ALAKANINVR AIAQGCSEYN ITVVVKREDC   540
 541 IKALRAVHSR FYLSRTTIAM GIIGPGLIGS TLLEQLRDQA STLKEEFNID LRVMGILGSK   600
 601 SMLLSDVGID LARWRELREE RGEVANMEKF VQHVHGNHFI PNTALVDCTA DSVIAGYYYD   660
 661 WLRKGIHVVT PNKKANSGPL DQYLKLRALQ RQSYTHYFYE ATVGAGLPIV STLRGLLETG   720
 721 DKILQIEGIF SGTLSYIFNN FKDGRAFSEV VSEAKEAGYT EPDPRDDLSG TDVARKVIIL   780
 781 ARESGLKLEL SNIPVESPVP EPLRACASAQ EFMQELPKFD QEFTKKQEDA ENAGEVLRYV   840
 841 GVVDVTNKKG VVELRRYKKD HPFAQLSGSD NIIAFTTRRY KDQPLIVRGP GAGAQVTAGG   900
 901 IFSDILRLAS YLGAPS
```

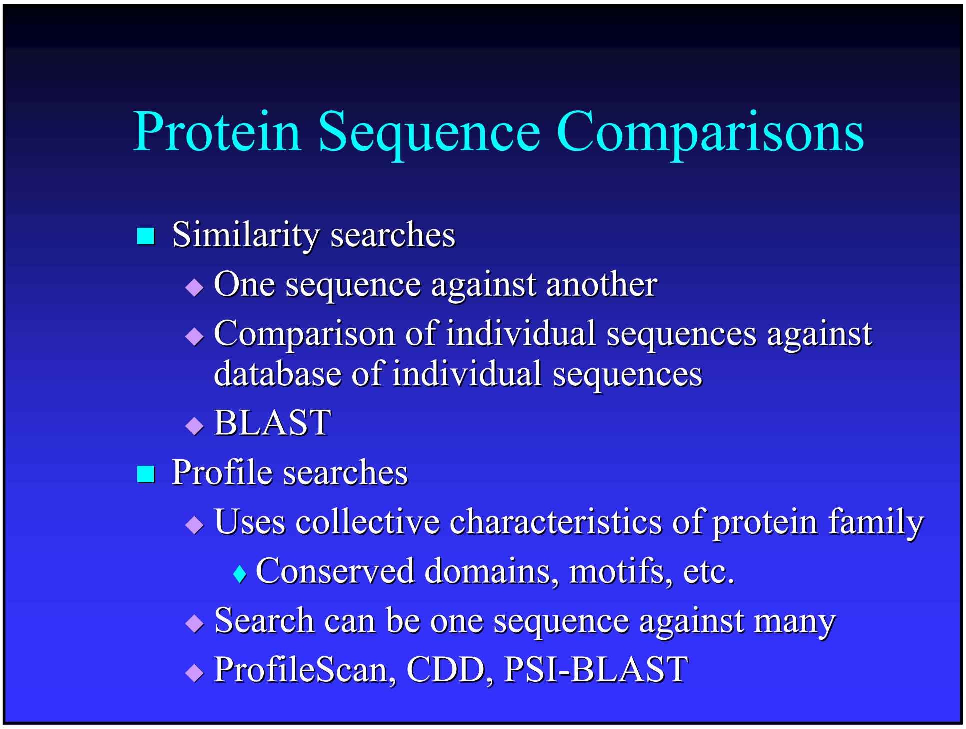
Molecular weight: 100394.54

Theoretical pI: 7.63

Pair-wise alignment of protein sequences

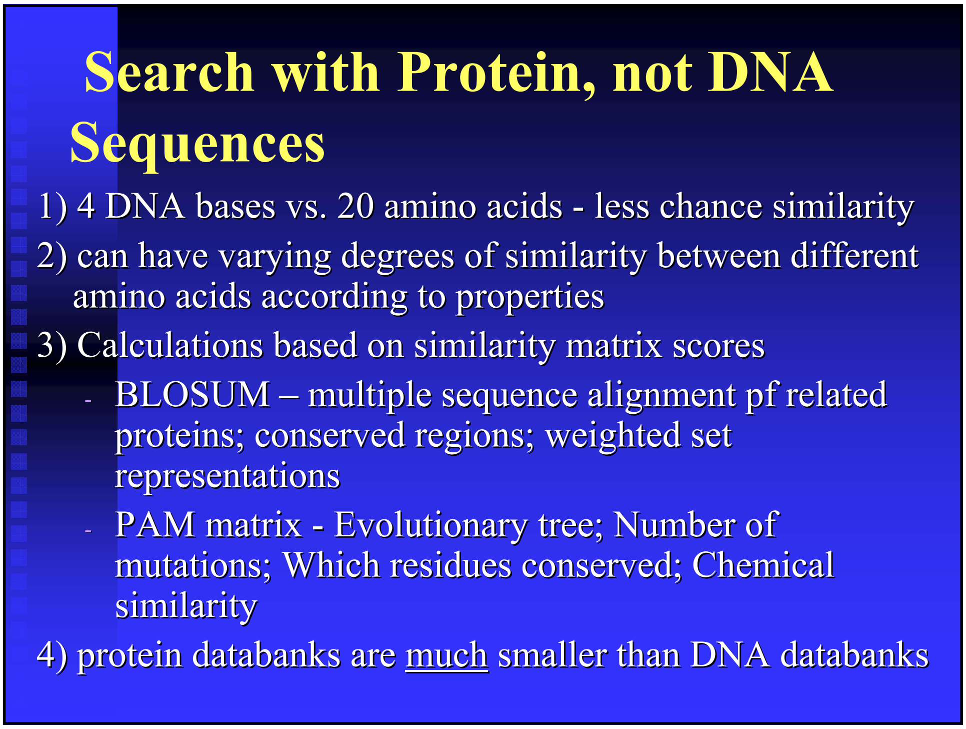
Why do Pairwise alignment searches?

- Are there other genes in database similar to yours?
- Have these other genes been well studied?
  - Leads to literature searches on these genes
- What is the function of these genes?
- Identify conserved motifs
  - Are they important to structure or function?
- Phylogenetic trees
  - Relatedness and evolution



Protein Sequence Comparisons

- Similarity searches
  - One sequence against another
  - Comparison of individual sequences against database of individual sequences
  - BLAST
- Profile searches
  - Uses collective characteristics of protein family
    - Conserved domains, motifs, etc.
  - Search can be one sequence against many
  - ProfileScan, CDD, PSI-BLAST



Search with Protein, not DNA Sequences

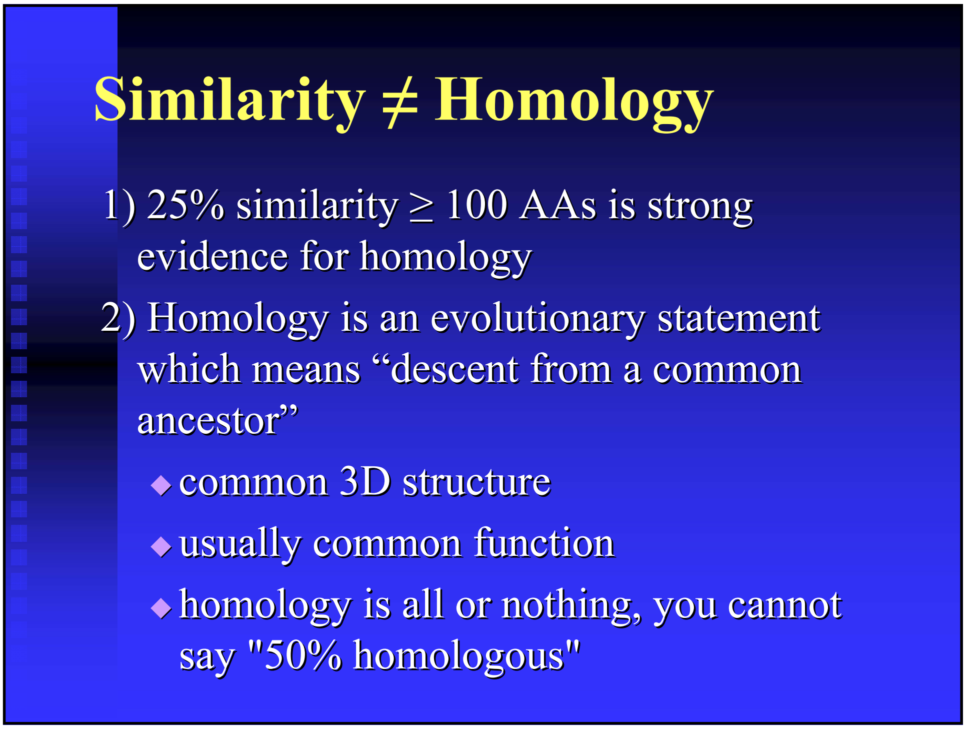
1) 4 DNA bases vs. 20 amino acids - less chance similarity

2) can have varying degrees of similarity between different amino acids according to properties

3) Calculations based on similarity matrix scores

- BLOSUM – multiple sequence alignment pf related proteins; conserved regions; weighted set representations
- PAM matrix - Evolutionary tree; Number of mutations; Which residues conserved; Chemical similarity

4) protein databanks are <u>much</u> smaller than DNA databanks

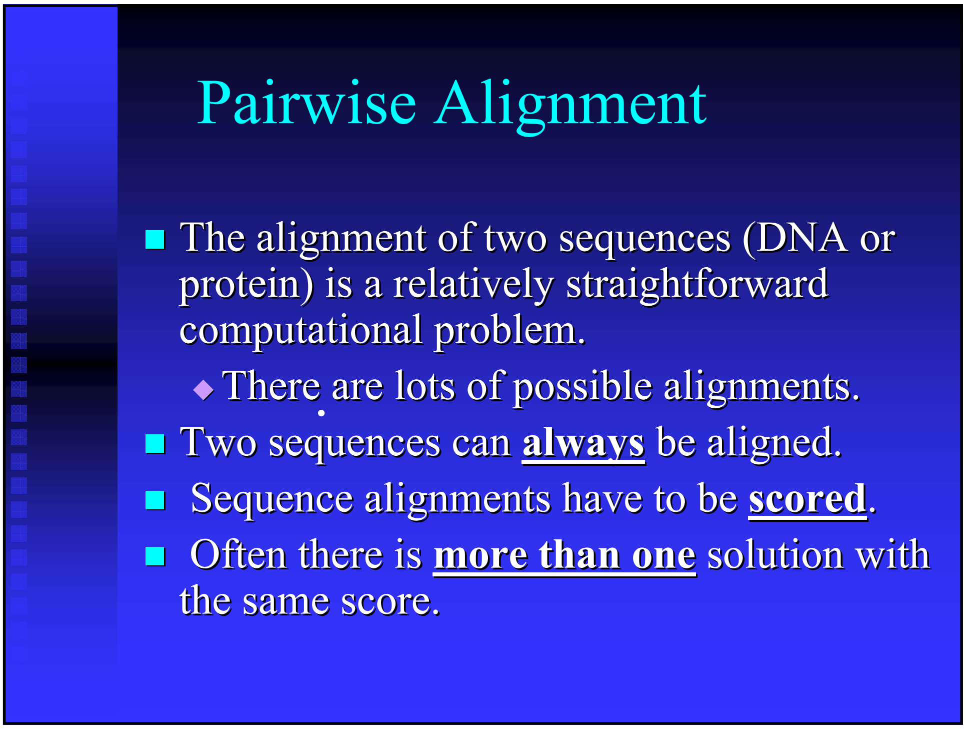


Similarity ≠ Homology

1) 25% similarity ≥ 100 AAs is strong evidence for homology

2) Homology is an evolutionary statement which means “descent from a common ancestor”

- common 3D structure
- usually common function
- homology is all or nothing, you cannot say "50% homologous"



Pairwise Alignment

- The alignment of two sequences (DNA or protein) is a relatively straightforward computational problem.
  - There are lots of possible alignments.
- Two sequences can **always** be aligned.
- Sequence alignments have to be **scored**.
- Often there is **more than one** solution with the same score.

Methods of Alignment

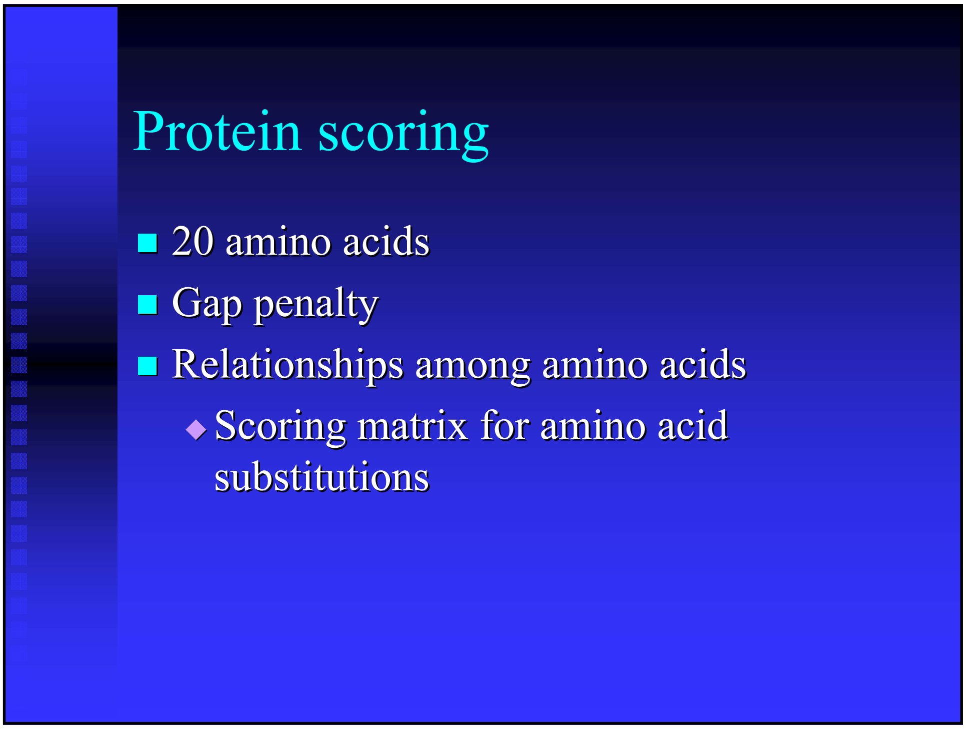
- By hand - slide sequences on two lines of a word processor
- Dot plot
  - with windows
- Rigorous mathematical approach
  - Dynamic programming (slow, optimal)
- Heuristic methods (fast, approximate)
  - BLAST and FASTA
    - Word matching and hash tables

DNA Scoring Systems
-very simple
-match or no match

```
Sequence 1   actaccagttcatttgatacttctcaaa
                      | |     |       ||
Sequence 2       taccattaccgtgttaactgaaaggacttaaagact
```

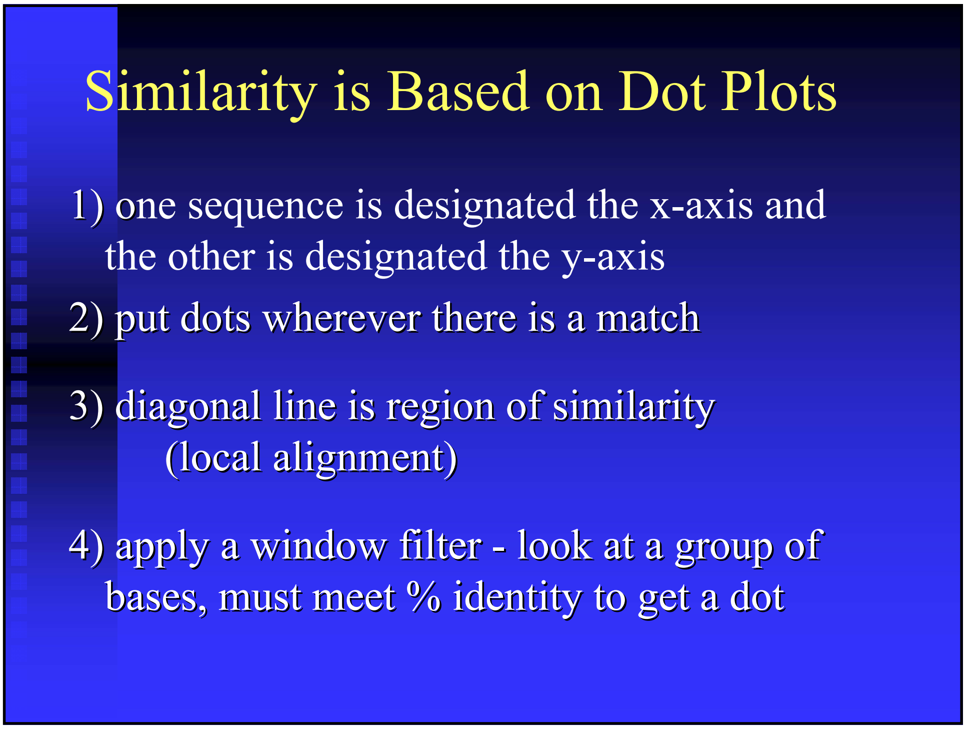
	A	G	C	T
A	**1**	0	0	0
G	0	**1**	0	0
C	0	0	**1**	0
T	0	0	0	**1**

Match: 1
Mismatch: 0
Score = 5



Protein scoring

- 20 amino acids
- Gap penalty
- Relationships among amino acids
  - Scoring matrix for amino acid substitutions



Similarity is Based on Dot Plots

1) one sequence is designated the x-axis and the other is designated the y-axis

2) put dots wherever there is a match

3) diagonal line is region of similarity
(local alignment)

4) apply a window filter - look at a group of bases, must meet % identity to get a dot

Dot Matrix method

- One sequence is designated the x-axis and the other is designated the y-axis
- A dot is created when the sequence elements corresponding to the x and y coordinates “match”.
- Diagonal lines within these plots indicate regions of similarity.

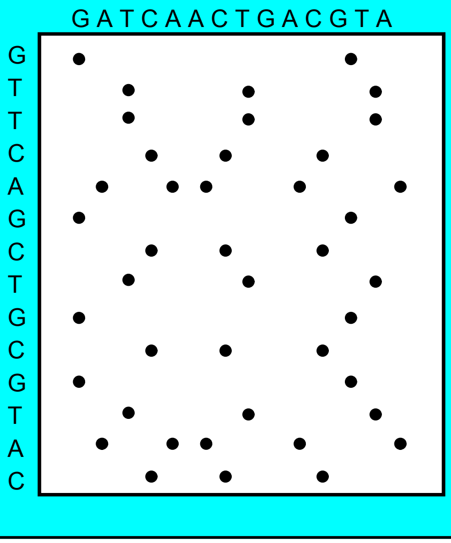
Simple Dot Matrix

	B	A	S	K	E	T	S	L	L	L
B	•									
A		•								
S			•							
E					•					
B	•									
A		•								
L								•	•	•

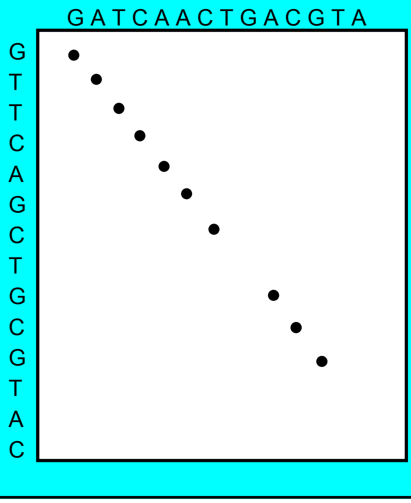
Characteristics of Dot Matrix

- All possible matches of residues between two sequences are found
- Reveal the presence of insertions/deletions and direct and inverted repeats
- Dot matrix is visible on the computer screen
- Limitation is that most dot matrix computer programs do not show an actual alignment.

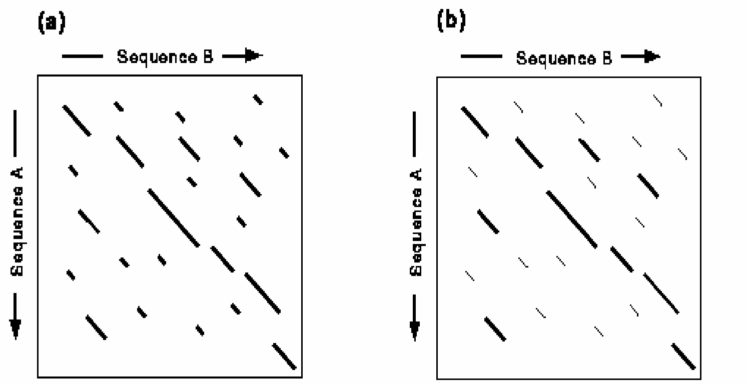
Simple Dot Plot



Dot plot filtered with 4 base window and 75% identity



FASTA Algorithm



Find runs of identitical words

Re-score using PAM matrix

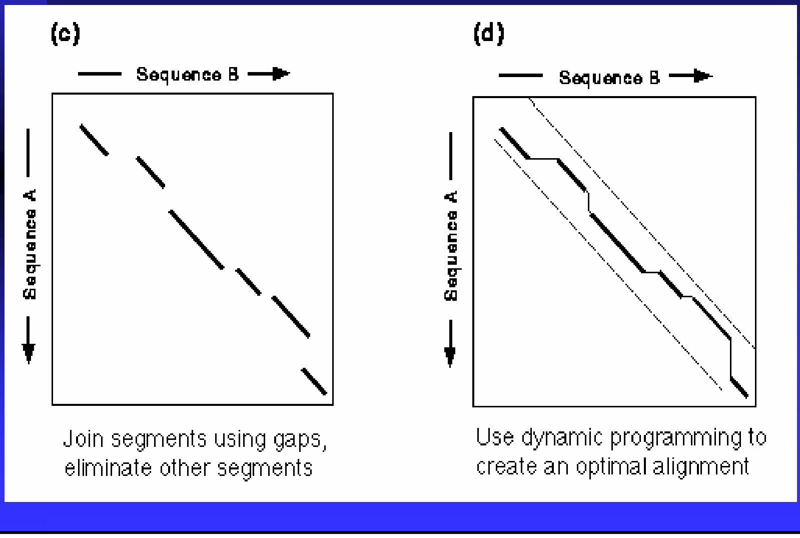
Keep top scoring segments

Makes Longest Diagonal

3) after all diagonals found, tries to join diagonals by adding gaps

4) computes alignments in regions of best diagonals

FASTA Alignments



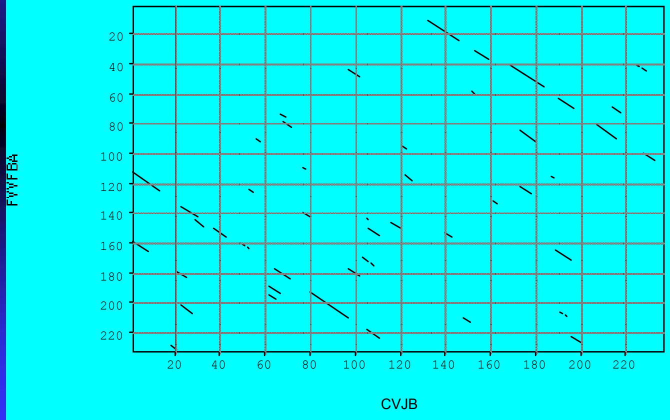
Join segments using gaps, eliminate other segments

Use dynamic programming to create an optimal alignment

Dot plot of real data

Window Size = 8
Min. % Score = 30
Hash Value = 2

Scoring Matrix: pam250 matrix



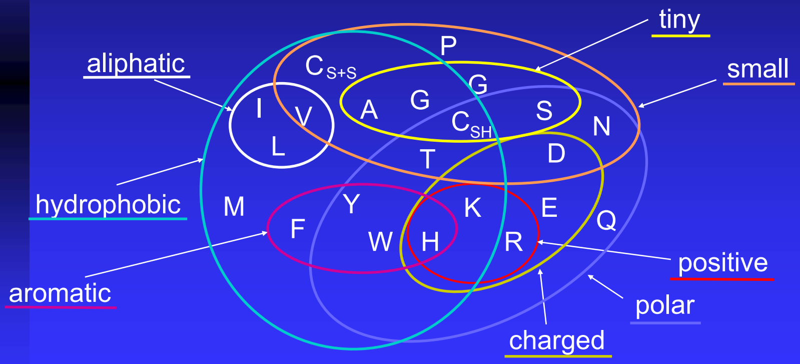
Amino acid scoring matrix

Protein Alignment Scoring Matrix Is Complex

- Conservation: What residues can substitute for another residue and not adversely affect the function of the protein?
  - Isoleucine and valine are both small and hydrophobic
  - Serine and threonine are both polar
  - Conserve charge, size, hydrophobicity, and other physicochemical factors
- Frequency:
  - How often does a particular residue occur
  - How ofter does it change? And to what other amino acid?

Protein Scoring Systems

- Amino acids have different biochemical and physical properties that influence their relative replaceability in evolution.

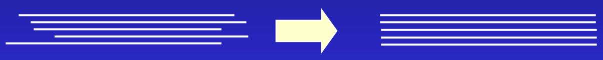


Scoring Matrix

- Important to understand scoring matrices
  - Play a role in all analyses involving sequence comparison
  - Assumptions are made
  - Which assumptions agree with what you want?
  - Choice of matrix (thus software) can strongly influence outcome

PAM (**P**ercent **A**ccepted **M**utations) matrices

- Derived from global alignments of **protein families** . Family members share at least 85% identity (Dayhoff *et al*., 1978).



- *Construc*tion of phylogenetic tree and ancestral sequences of each protein family

- Computation of number of replacements for each pair of amino acids

•The number following the matirx, PAM30 or PAM100 refer to eevolutionary distance; the greater the number, the greater the distance.

PAM (Percent Accepted Mutations) matrices

- The numbers of replacements were used to compute a so-called PAM-1 matrix.

- The PAM-1 matrix reflects an average change of 1% of all amino acid positions, ie. roughly 1% divergence. PAM matrices for larger evolutionary distances can be extrapolated from the PAM-1 matrix.

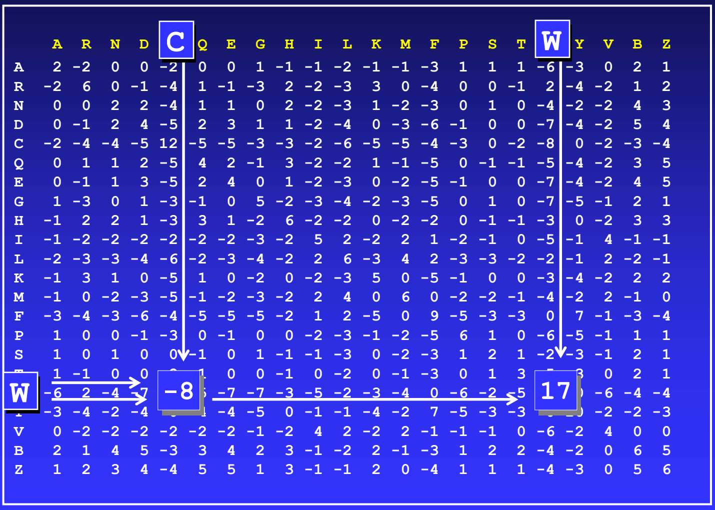
- PAM250 = 250 mutations per 100 residues.

- Greater numbers mean bigger evolutionary distance

- Analysis documented 1572 changes in 71 groups of protein

- High similarity within original sequence set, represents substitution pattern expected over short evolutionary distance

PAM 250



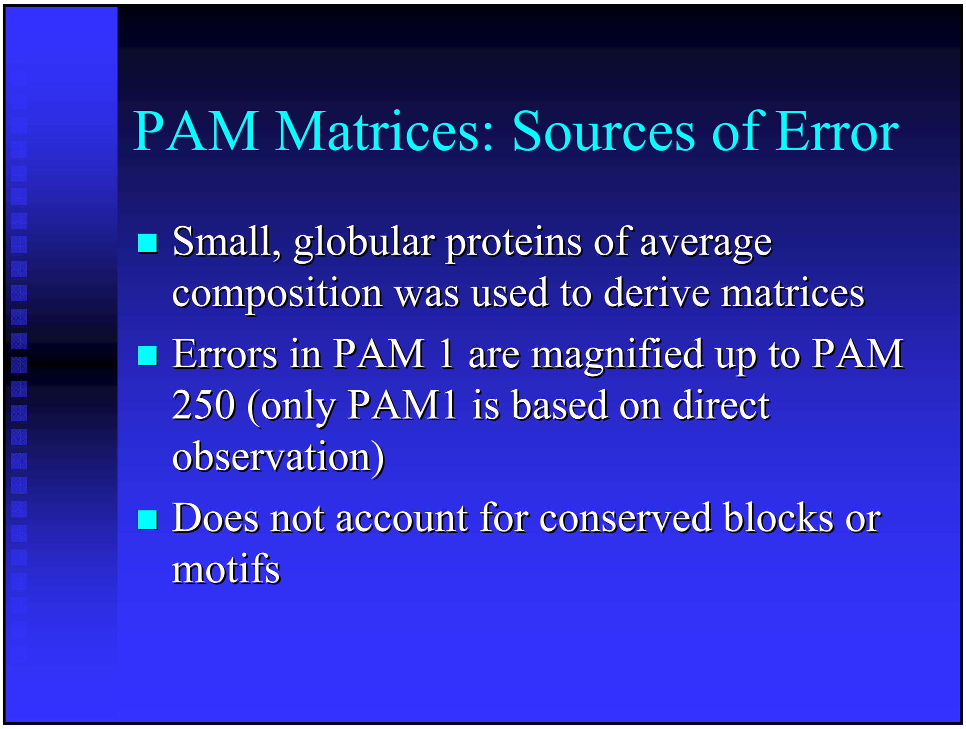
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z
A	2	-2	0	0	-2	0	0	1	-1	-1	-2	-1	-1	-3	1	1	1	-6	-3	0	2	1
R	-2	6	0	-1	-4	1	-1	-3	2	-2	-3	3	0	-4	0	0	-1	2	-4	-2	1	2
N	0	0	2	2	-4	1	1	0	2	-2	-3	1	-2	-3	0	1	0	-4	-2	-2	4	3
D	0	-1	2	4	-5	2	3	1	1	-2	-4	0	-3	-6	-1	0	0	-7	-4	-2	5	4
C	-2	-4	-4	-5	12	-5	-5	-3	-3	-2	-6	-5	-5	-4	-3	0	-2	-8	0	-2	-3	-4
Q	0	1	1	2	-5	4	2	-1	3	-2	-2	1	-1	-5	0	-1	-1	-5	-4	-2	3	5
E	0	-1	1	3	-5	2	4	0	1	-2	-3	0	-2	-5	-1	0	0	-7	-4	-2	4	5
G	1	-3	0	1	-3	-1	0	5	-2	-3	-4	-2	-3	-5	0	1	0	-7	-5	-1	2	1
H	-1	2	2	1	-3	3	1	-2	6	-2	-2	0	-2	-2	0	-1	-1	-3	0	-2	3	3
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5	2	-2	2	1	-2	-1	0	-5	-1	4	-1	-1
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6	-3	4	2	-3	-3	-2	-2	-1	2	-2	-1
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5	0	-5	-1	0	0	-3	-4	-2	2	2
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6	0	-2	-2	-1	-4	-2	2	-1	0
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9	-5	-3	-3	0	7	-1	-3	-4
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6	1	0	-6	-5	-1	1	1
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2	1	-2	-3	-1	2	1
T	1	-1	0	0	[illegible]	-1	0	0	-1	0	-2	0	-1	-3	0	1	3	-5	-3	0	2	1
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17	0	-6	-4	-4
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	-2	-2	-3
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4	0	0
B	2	1	4	5	-3	3	4	2	3	-1	-2	2	-1	-3	1	2	2	-4	-2	0	6	5
Z	1	2	3	4	-4	5	5	1	3	-1	-1	2	0	-4	1	1	1	-4	-3	0	5	6

PAM Matrices

- Short evolutionary distance
  - Change in function unlikely
- Point Accepted Mutation (PAM)
  - The new side chain must function the same way as old one (“acceptance”)
  - On average, 1 PAM corresponds to 1 amino acid change per 100 residues
  - 1 PAM ~1% divergence
  - Extrapolates to predict patterns at longer evolutionary distances

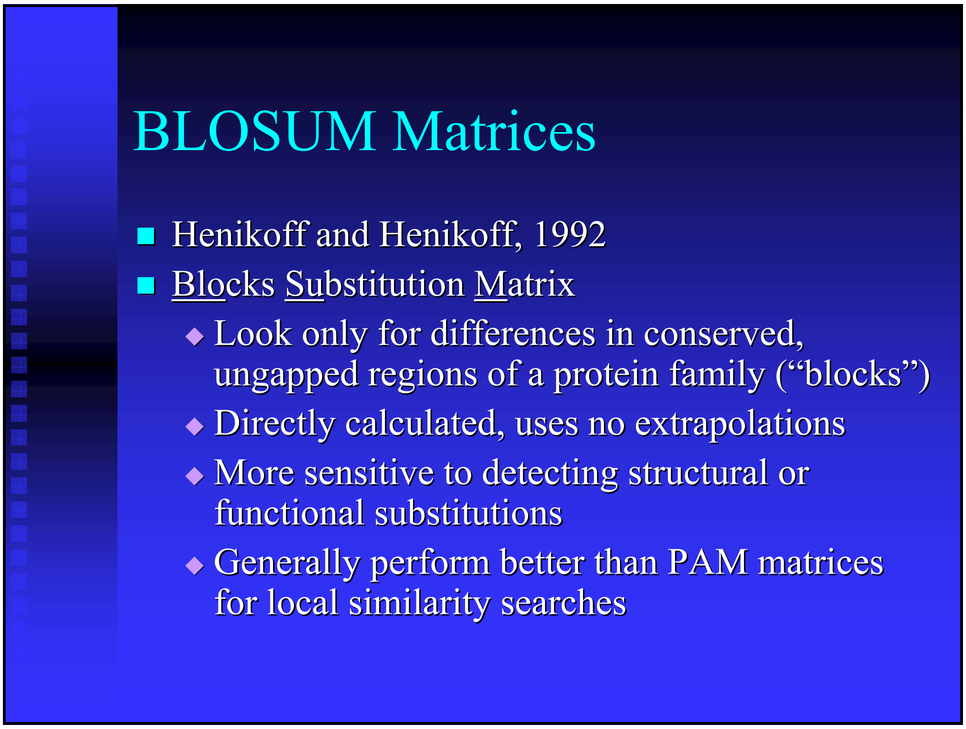
PAM Matrices: Assumptions

- All sites assumed to be equally mutable
- Replacement of amino acids is independent of previous mutations at the same position
- Replacement is independent of surrounding residues
- Forces responsible for sequence evolution over shorter time spans are the same as those over longer time spans



PAM Matrices: Sources of Error

- Small, globular proteins of average composition was used to derive matrices
- Errors in PAM 1 are magnified up to PAM 250 (only PAM1 is based on direct observation)
- Does not account for conserved blocks or motifs

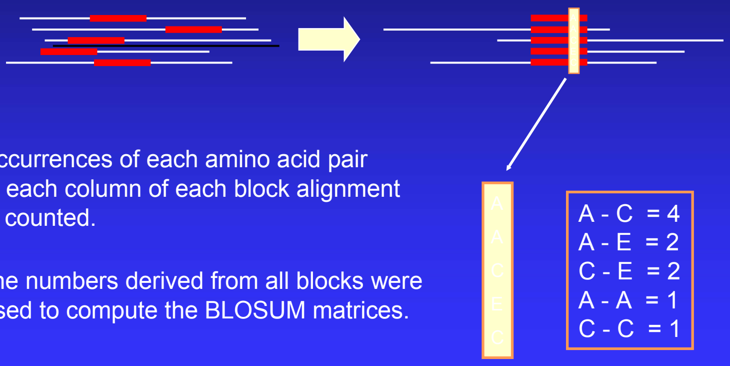


BLOSUM Matrices

- Henikoff and Henikoff, 1992
- Blocks Substitution Matrix
  - Look only for differences in conserved, ungapped regions of a protein family (“blocks”)
  - Directly calculated, uses no extrapolations
  - More sensitive to detecting structural or functional substitutions
  - Generally perform better than PAM matrices for local similarity searches

BLOSUM (**Blo**cks **Su**bstitution **M**atrix)

- Derived from alignments of domains of **distantly** related proteins (Henikoff & Henikoff,1992).



A - C = 4
A - E = 2
C - E = 2
A - A = 1
C - C = 1

- Occurrences of each amino acid pair in each column of each block alignment is counted.

- The numbers derived from all blocks were used to compute the BLOSUM matrices.

BLOSUM (**Blo**cks **Su**bstitution **M**atrix)

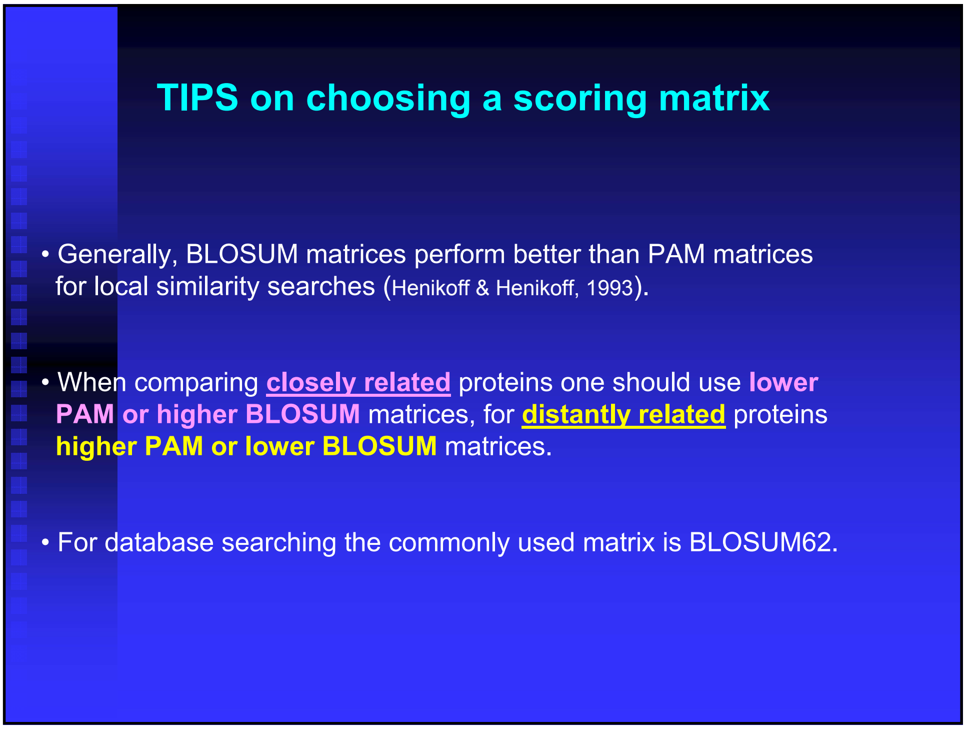
- Sequences within blocks are clustered according to their level of identity.

- Clusters are counted as a single sequence.

- Different BLOSUM matrices differ in the percentage of sequence identity used in clustering.

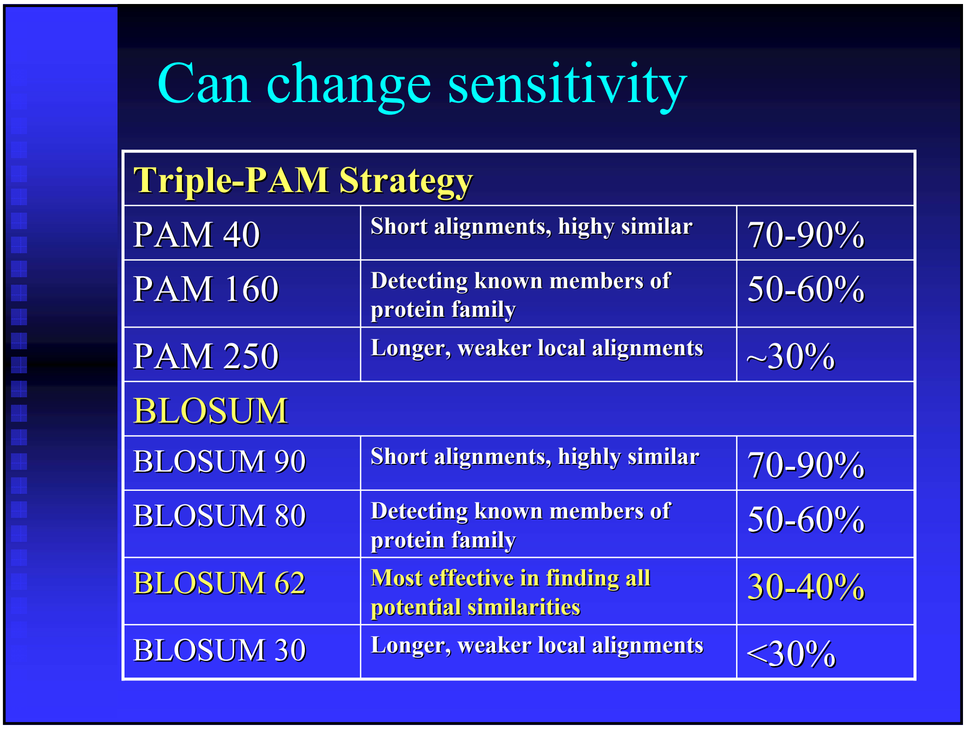
- The number in the matrix name (e.g. 62 in BLOSUM62) refers to the percentage of sequence identity used to build the matrix.

- Greater numbers mean smaller evolutionary distance.



TIPS on choosing a scoring matrix

- Generally, BLOSUM matrices perform better than PAM matrices for local similarity searches (Henikoff & Henikoff, 1993).
- When comparing **closely related** proteins one should use **lower PAM or higher BLOSUM** matrices, for **distantly related** proteins **higher PAM or lower BLOSUM** matrices.
- For database searching the commonly used matrix is BLOSUM62.

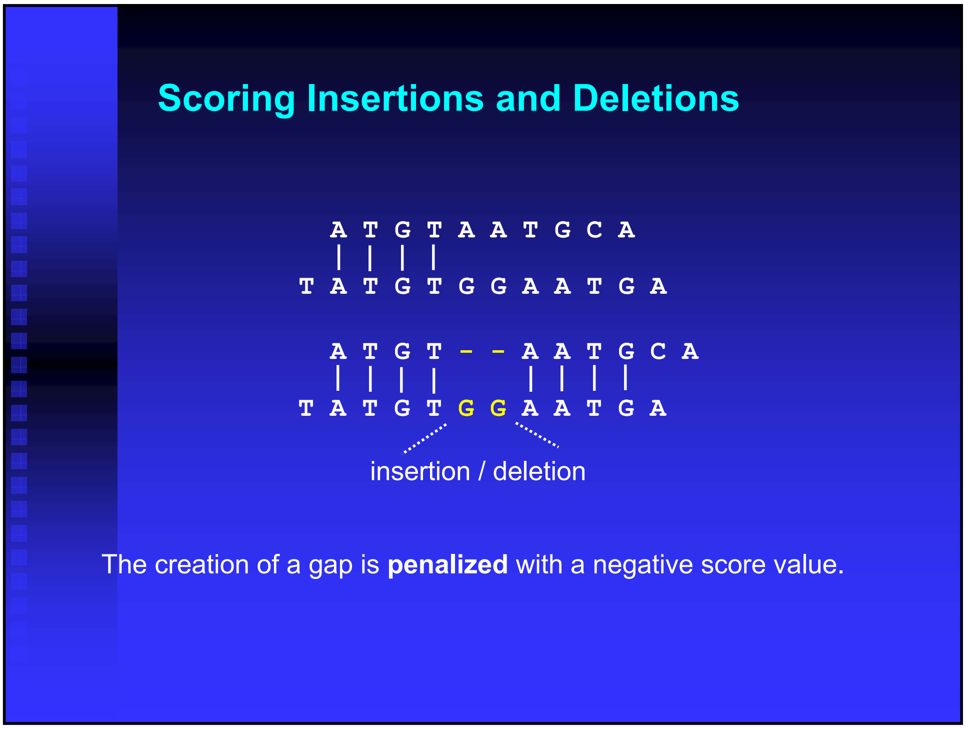


Can change sensitivity

Triple-PAM Strategy		
PAM 40	**Short alignments, highy similar**	70-90%
PAM 160	**Detecting known members of protein family**	50-60%
PAM 250	**Longer, weaker local alignments**	~30%
BLOSUM		
BLOSUM 90	**Short alignments, highly similar**	70-90%
BLOSUM 80	**Detecting known members of protein family**	50-60%
BLOSUM 62	**Most effective in finding all potential similarities**	30-40%
BLOSUM 30	**Longer, weaker local alignments**	<30%

No single matrix is the complete answer for all sequence comparisons

Gap penalties



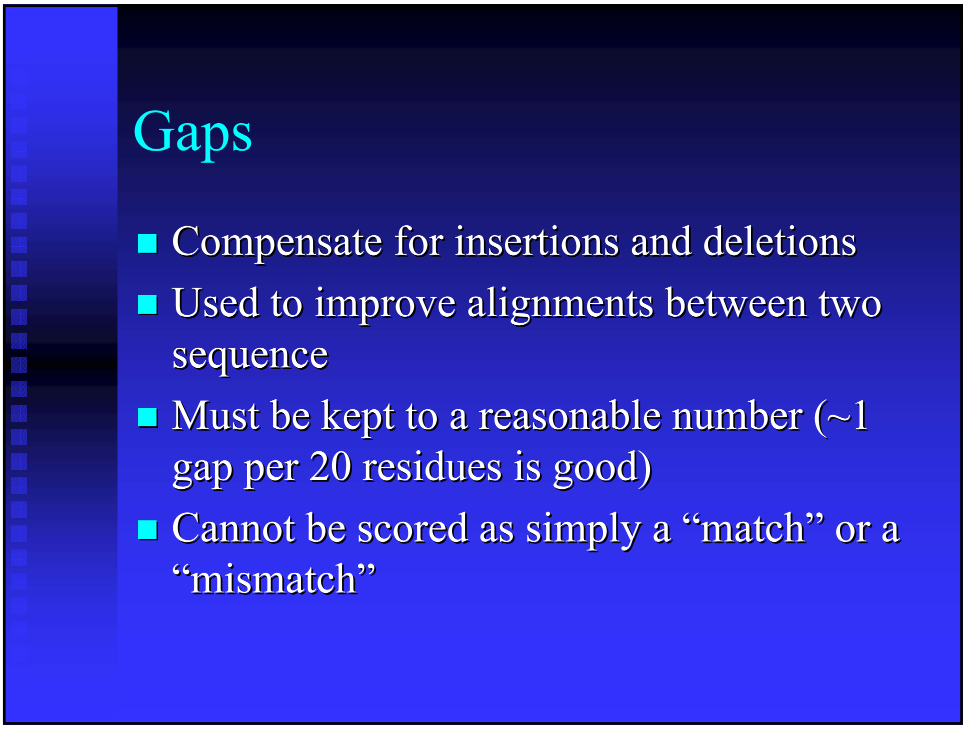
Scoring Insertions and Deletions

```
  A T G T A A T G C A
  | | | |
T A T G T G G A A T G A

  A T G T - - A A T G C A
  | | | |     | | | |
T A T G T G G A A T G A
```

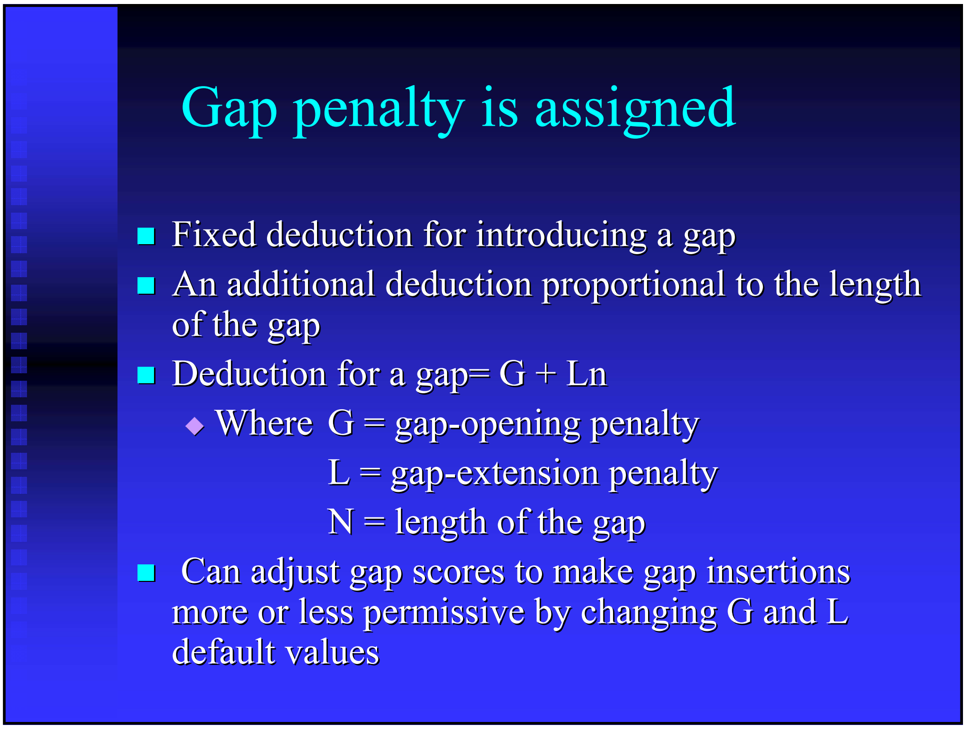
insertion / deletion

The creation of a gap is **penalized** with a negative score value.



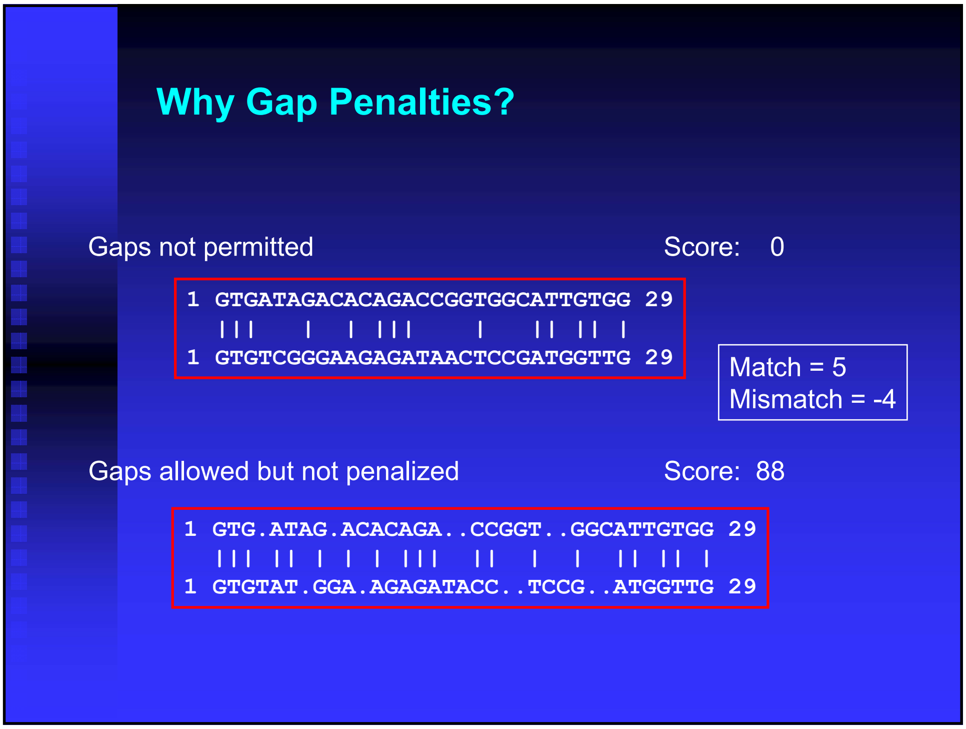
Gaps

- Compensate for insertions and deletions
- Used to improve alignments between two sequence
- Must be kept to a reasonable number (~1 gap per 20 residues is good)
- Cannot be scored as simply a “match” or a “mismatch”



Gap penalty is assigned

- Fixed deduction for introducing a gap
- An additional deduction proportional to the length of the gap
- Deduction for a gap= G + Ln
  - Where G = gap-opening penalty
    L = gap-extension penalty
    N = length of the gap
- Can adjust gap scores to make gap insertions more or less permissive by changing G and L default values



Why Gap Penalties?

Gaps not permitted Score: 0

```
1 GTGATAGACACAGACCGGTGGCATTGTGG 29
  |||   |  | |||    |   || || |
1 GTGTCGGGAAGAGATAACTCCGATGGTTG 29
```

Match = 5
Mismatch = -4

Gaps allowed but not penalized Score: 88

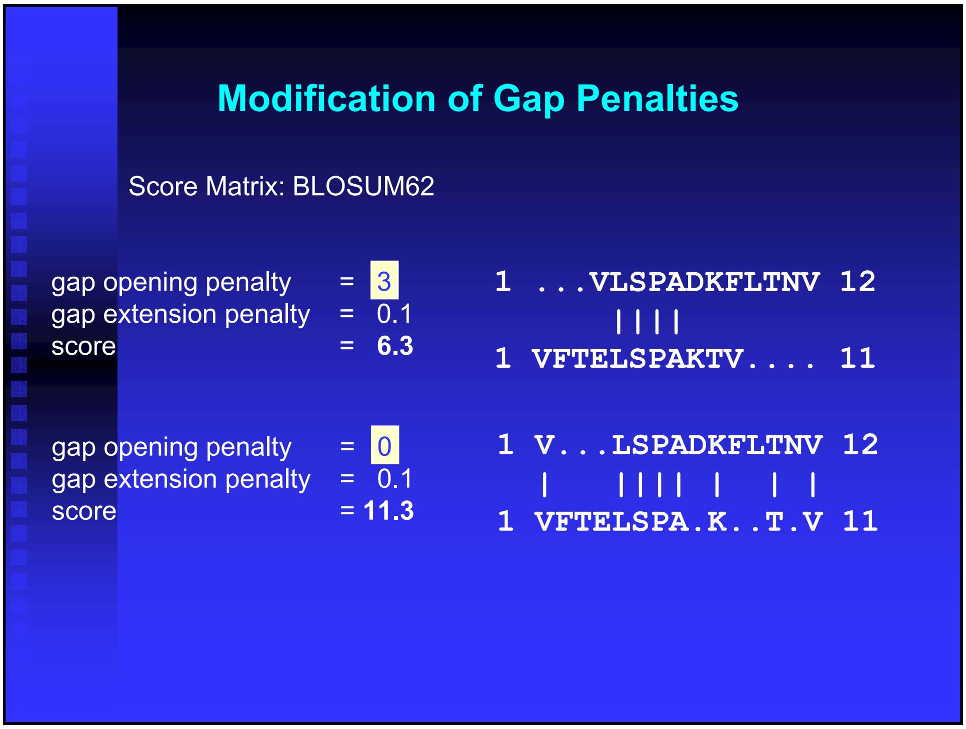
```
1 GTG.ATAG.ACACAGA..CCGGT..GGCATTGTGG 29
  ||| || | | | |||  ||  |  |  || || |
1 GTGTAT.GGA.AGAGATACC..TCCG..ATGGTTG 29
```

Why Gap Penalties?

- The optimal alignment of two similar sequences is usually that which
  - **maximizes** the number of matches and
  - **minimizes** the number of gaps.
  - There is a tradeoff between these two
    - adding gaps reduces mismatches
- Permitting the insertion of arbitrarily many gaps can lead to high scoring alignments of **non-homologous** sequences.
- Penalizing gaps forces alignments to have relatively few gaps.

Gap Penalties

- How to balance gaps with mismatches?
- Gaps must get a steep penalty, or else you'll end up with nonsense alignments.
- In real sequences, muti-base (or amino acid) gaps are quit common
  - genetic insertion/deletion events
- "Affine" gap penalties give a big penalty for each new gap, but a much smaller "gap extension" penalty.



Modification of Gap Penalties

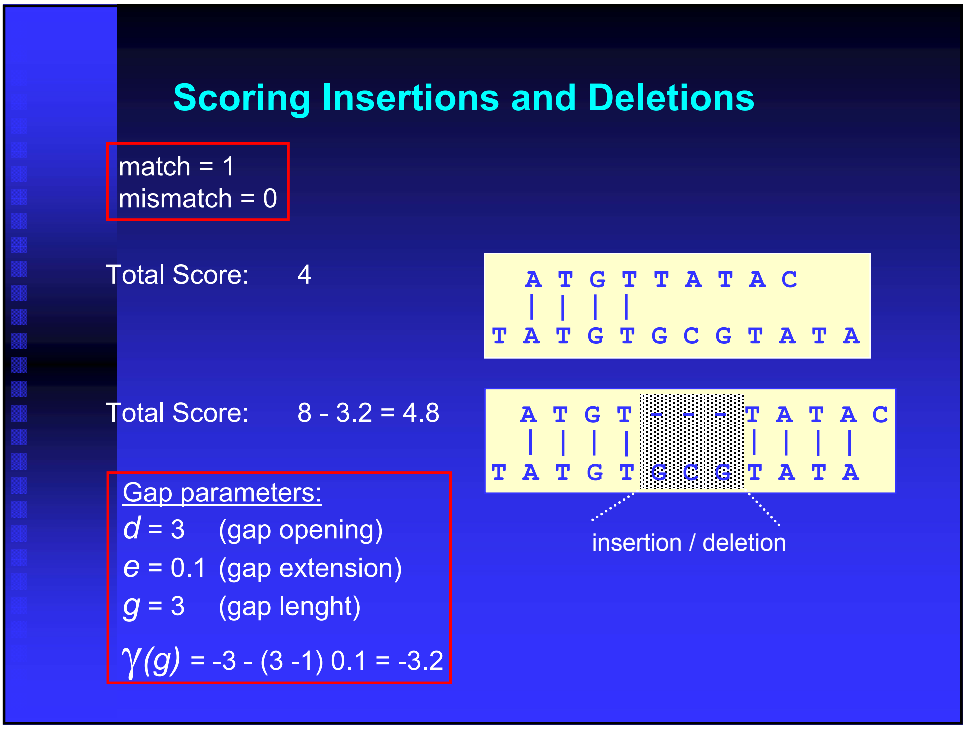
Score Matrix: BLOSUM62

gap opening penalty = 3
gap extension penalty = 0.1
score = **6.3**

```
1 ...VLSPADKFLTNV 12
      ||||
1 VFTELSPAKTV.... 11
```

gap opening penalty = 0
gap extension penalty = 0.1
score = **11.3**

```
1 V...LSPADKFLTNV 12
  |   |||| |  | |
1 VFTELSPA.K..T.V 11
```



Scoring Insertions and Deletions

match = 1
mismatch = 0

Total Score: 4

```
  A T G T T A T A C
  | | | |
T A T G T G C G T A T A
```

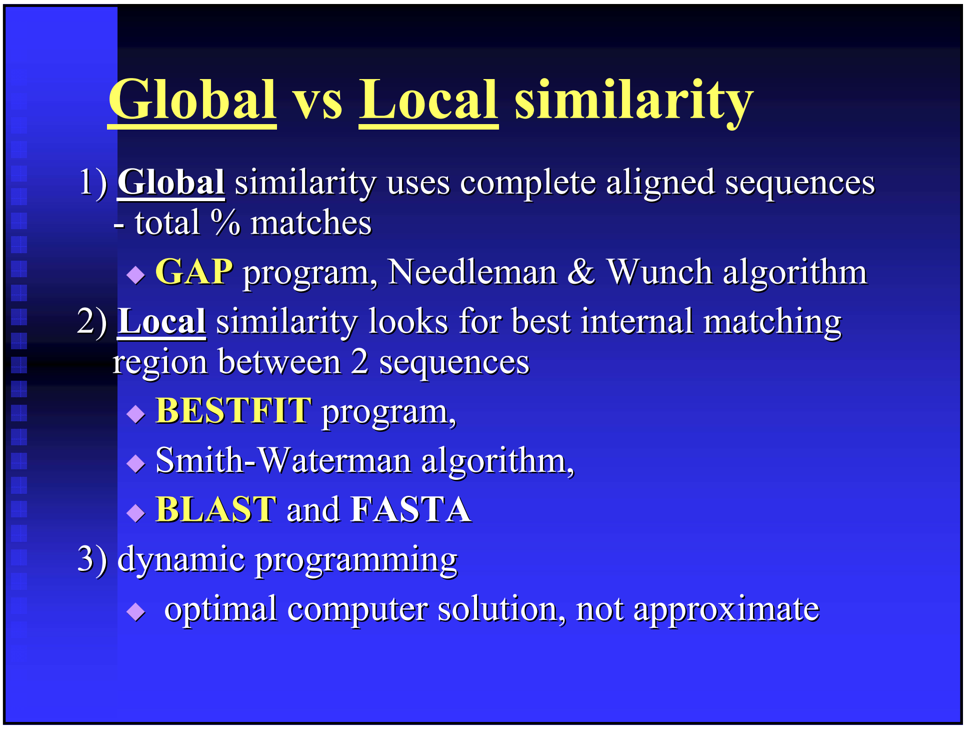
Total Score: 8 - 3.2 = 4.8

```
  A T G T - - - T A T A C
  | | | |       | | | |
T A T G T G C G T A T A
```

insertion / deletion

Gap parameters:
d = 3 (gap opening)
e = 0.1 (gap extension)
g = 3 (gap lenght)

$\gamma(g)$ = -3 - (3 -1) 0.1 = -3.2



Global vs Local similarity

1) **Global** similarity uses complete aligned sequences - total % matches

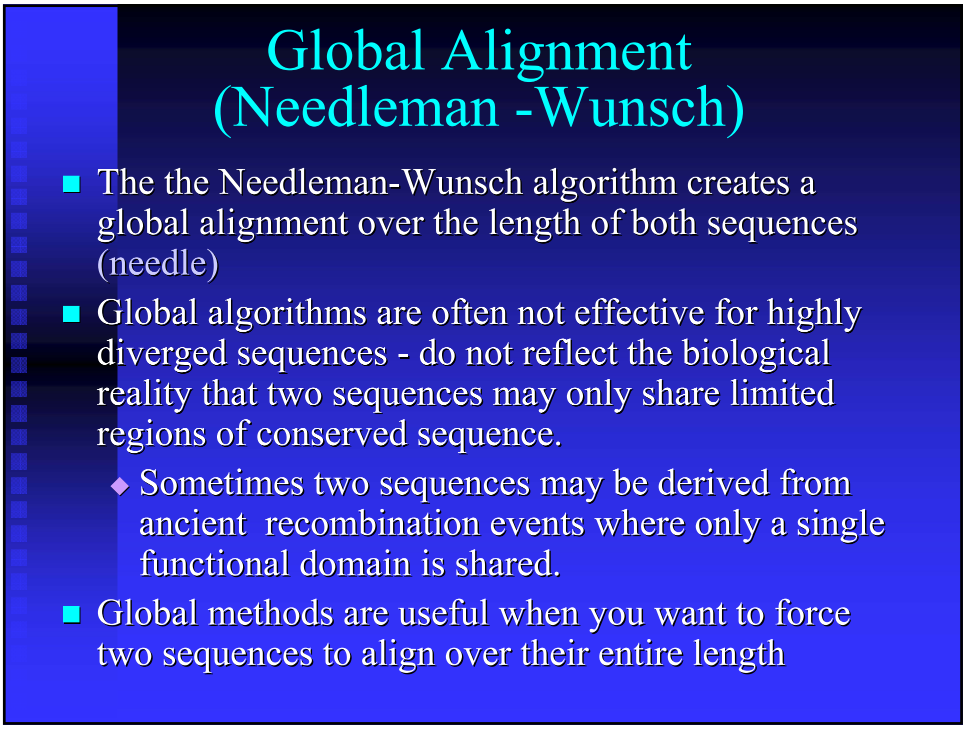
- **GAP** program, Needleman & Wunch algorithm

2) **Local** similarity looks for best internal matching region between 2 sequences

- **BESTFIT** program,
- Smith-Waterman algorithm,
- **BLAST** and **FASTA**

3) dynamic programming

- optimal computer solution, not approximate



Global Alignment (Needleman -Wunsch)

- The the Needleman-Wunsch algorithm creates a global alignment over the length of both sequences (needle)
- Global algorithms are often not effective for highly diverged sequences - do not reflect the biological reality that two sequences may only share limited regions of conserved sequence.
  - Sometimes two sequences may be derived from ancient recombination events where only a single functional domain is shared.
- Global methods are useful when you want to force two sequences to align over their entire length

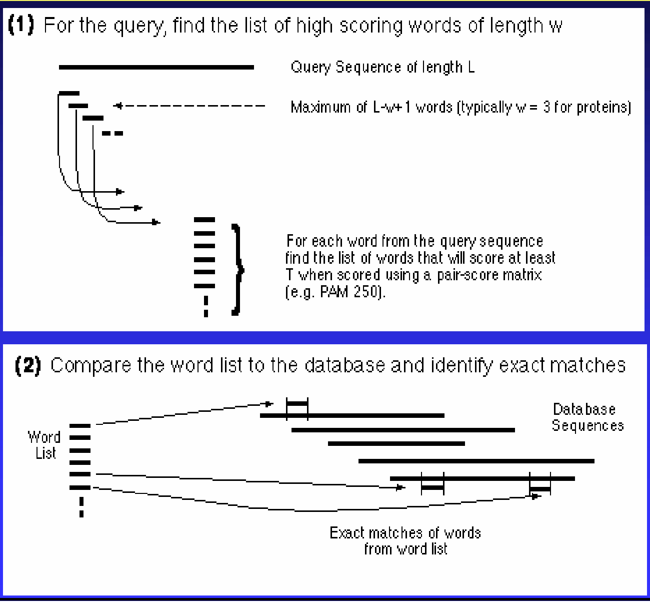
Local Alignment (Smith-Waterman)

- Local alignment
  - Identify the most similar sub-region shared between two sequences
  - Smith-Waterman

Scoring Similarity

1) Can only score aligned sequences
2) DNA is usually scored as identical or not
3) Amino acids have varying degrees of similarity
   - a. # of mutations to convert one to another
   - b. chemical similarity
   - c. observed mutation frequencies
4) Modified scoring for gaps - single vs. multiple base gaps (gap extension)
5) PAM matrix calculated from observed mutations in protein families
6) BLOSUM matrix calculated from changes in conserved blocks of amino acid sequenc

BLAST Algorithm



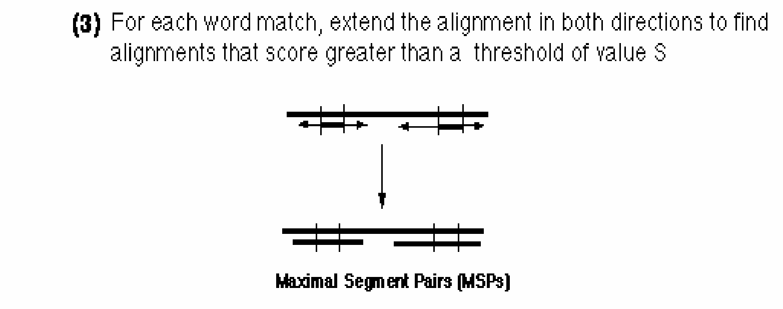
(1) For the query, find the list of high scoring words of length w

Query Sequence of length L
Maximum of L-w+1 words (typically w = 3 for proteins)
For each word from the query sequence find the list of words that will score at least T when scored using a pair-score matrix (e.g. PAM 250).

(2) Compare the word list to the database and identify exact matches

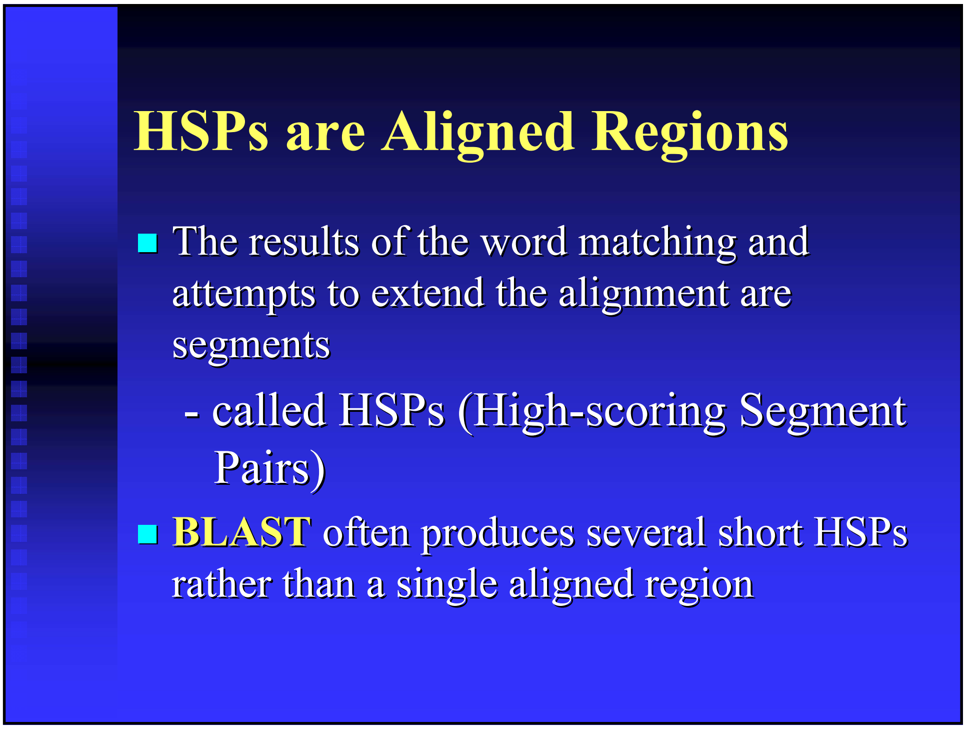
Word List
Database Sequences
Exact matches of words from word list

Extend hits one base at a time



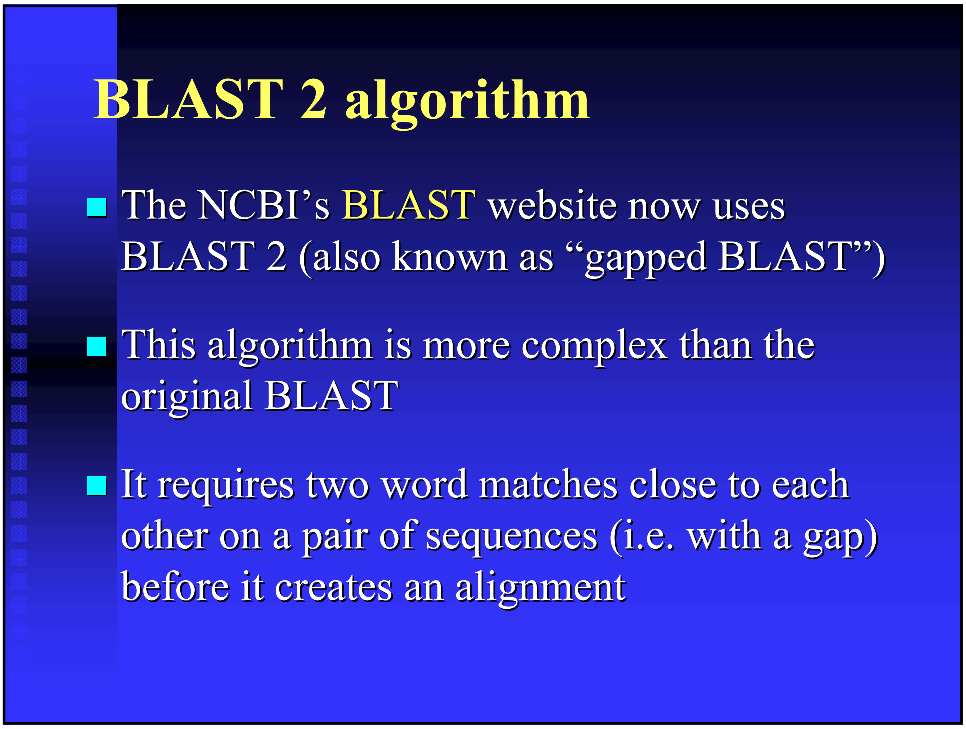
(3) For each word match, extend the alignment in both directions to find alignments that score greater than a threshold of value S

Figure from Barton, G.J. Protein Sequence Alignment and Database Scanning (University of Oxford, Laboratory of Molecular Biophysics)



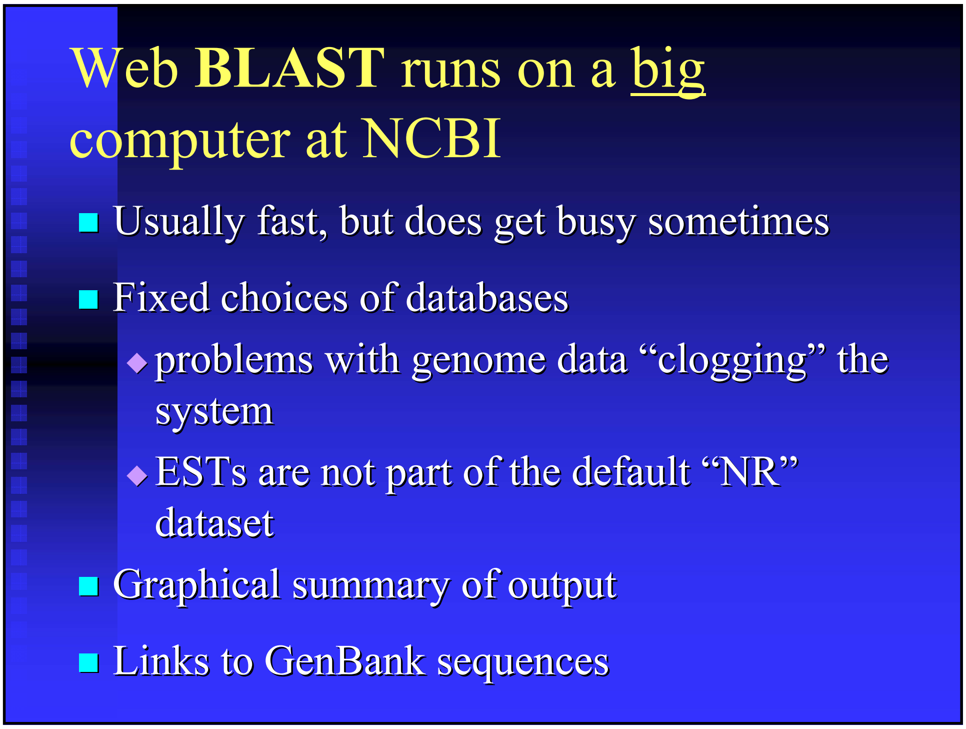
HSPs are Aligned Regions

- The results of the word matching and attempts to extend the alignment are segments
  - called HSPs (High-scoring Segment Pairs)
- **BLAST** often produces several short HSPs rather than a single aligned region



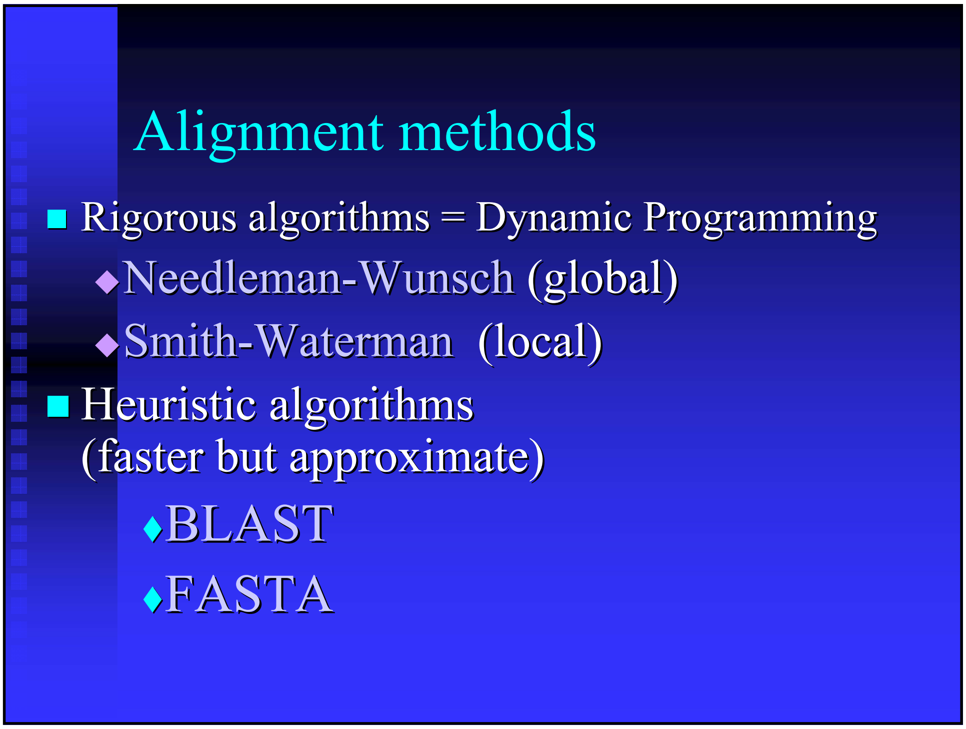
BLAST 2 algorithm

- The NCBI’s BLAST website now uses BLAST 2 (also known as “gapped BLAST”)
- This algorithm is more complex than the original BLAST
- It requires two word matches close to each other on a pair of sequences (i.e. with a gap) before it creates an alignment



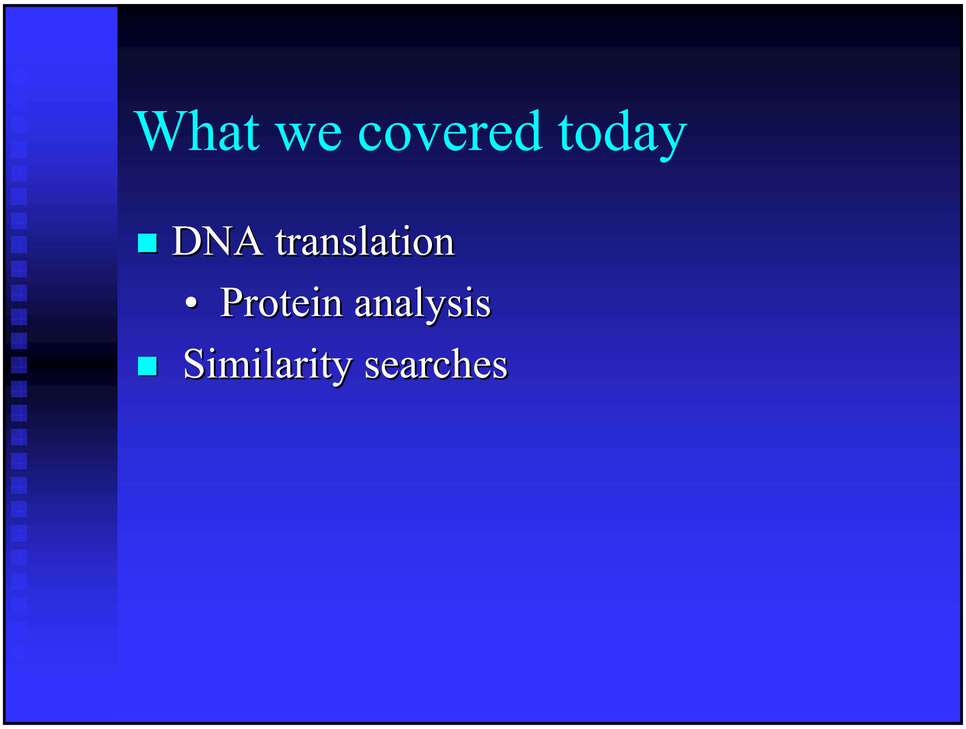
Web **BLAST** runs on a big computer at NCBI

- Usually fast, but does get busy sometimes
- Fixed choices of databases
  - problems with genome data “clogging” the system
  - ESTs are not part of the default “NR” dataset
- Graphical summary of output
- Links to GenBank sequences



Alignment methods

- Rigorous algorithms = Dynamic Programming
  - Needleman-Wunsch (global)
  - Smith-Waterman (local)
- Heuristic algorithms (faster but approximate)
  - BLAST
  - FASTA



What we covered today

- DNA translation
    - Protein analysis
- Similarity searches